Presidenza del Consiglio dei Ministri

UFFICIO DEL BILANCIO E PER IL RISCONTRO
DI REGOLARITA' AMMINISTRATIVO-CONTABILE
Servizio 5 - Riscontro atti organizzativi e atti relativi alle spese di personale

AL DIPARTIMENTO PER IL
COORDINAMENTO AMMINISTRATIVO
Via della Mercede, 9
00187 - ROMA

OGGETTO: Approvazione del progetto esecutivo Human Technopole.

Si restituisce, munito del visto di registrazione della Corte dei Conti, il provvedimento in oggetto.

IL COORDINATORE DEL SERVIZIO
(Dott. Gianfranco Sorchetti)

Il Presidente del Consiglio dei Ministri

VISTO il decreto legge 30 settembre 2003, n. 269, convertito, con modificazioni, dalla legge 24 novembre 2003, n. 326, recante "Diposizioni urgenti per favorire lo sviluppo e per la correzione dell'andamento dei conti pubblici", e in particolare l'articolo 4, che istituisce la fondazione denominata Istituto Italiano di Tecnologia;

VISTO lo statuto dell'Istituto Italiano di Tecnologia, e in particolare l'articolo 3, che attribuisce alla fondazione, tra l'altro, lo scopo di promuovere l'eccellenza scientifica e tecnologica sia in forma diretta, sia in forma indiretta, facendo leva su collaborazioni a rete con laboratori e gruppi di eccellenza nazionali e internazionali; di sviluppare, al suo interno e nell'ambito del sistema nazionale della ricerca, la cultura della condivisione e del valore dei risultati a fini produttivi e sociali; di promuove l'integrazione tra aree di ricerca e l'interazione tra ricerca di base e applicata, nonché lo sviluppo sperimentale;

VISTO il decreto legge 25 novembre 2015, n. 185, convertito, con modificazioni, dalla legge 22 gennaio 2016, n. 9, recante "Misure urgenti per interventi nel territorio", e in particolare l'articolo 5, comma 2, che stabilisce che, nell'ambito delle iniziative per la valorizzazione dell'area utilizzata per l'Expo, "*è attribuito all'Istituto Italiano di Tecnologia (IIT) un primo contributo dell'importo di 80 milioni di euro per l'anno 2015 per la realizzazione di un progetto scientifico e di ricerca, sentiti gli enti territoriali e le principali istituzioni scientifiche interessate, da attuarsi anche utilizzando parte delle aree in uso a Expo S.p.a. ove necessario previo loro adattamento. IIT elabora un progetto esecutivo che è approvato con decreto del presidente del Consiglio dei ministri, su proposta del Ministro dell'economia e delle finanze*";

CONSIDERATO che l'Istituto Italiano di Tecnologia ha predisposto il menzionato progetto, denominato *Human Technopole*, multidisciplinare e integrato, in tema di salute, genomica e *data science*, che prevede un primo periodo di avvio, con una spesa prevista di 79,9 milioni di euro, finalizzata alla predisposizione dell'infrastruttura logistica, tecnologica e organizzativa, in coerenza con il citato art. 5, comma 2, del decreto legge n. 185/2015.

CONSIDERATO che sul progetto sono stati sentiti il comune di Milano, la regione Lombardia, il Politecnico di Milano, l'Università degli Studi di Milano, l'Università degli Studi Milano Bicocca, gli istituti di ricerca clinica e ospedaliera di Milano, la Fondazione *Edmund Mach* di Trento, la Fondazione ISI – Istituto per l'interscambio scientifico di Torino, il Consorzio interuniversitario CINECA di Bologna e il Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria – CREA;

CONSIDERATO che l'Università statale di Milano, con determinazione adottata dai propri organi di gestione, ha inteso formulare una specifica manifestazione d'interesse per l'insediamento in Arexpo del nuovo progetto di *Campus* scientifico universitario denominato "*Science of Citizens*" e che ciò costituisce una ulteriore opportunità di sviluppo per i progetti di ricerca di cui al presente decreto;

CONSIDERATO, altresì, che il progetto richiamato è stato oggetto di uno specifico processo di valutazione internazionale anonima, promosso dal MIUR e conclusosi positivamente con nota del 1° luglio 2016;

RITENUTO, pertanto, di poter procedere all'approvazione del progetto predisposto dall'Istituto Italiano di Tecnologia;

RITENUTO necessario, inoltre, definire l'impianto operativo che consentirà di avviare tempestivamente la prima fase di realizzazione del progetto scientifico, identificando le istituzioni coinvolte e le rispettive competenze, fino alla definizione normativa entro centoventi giorni dalla data del presente decreto, delle modalità di istituzione di un apposito ente, dotato di adeguato finanziamento, da costituire entro ventiquattro mesi dalla data del presente decreto, una volta compiute le attività di cui all'articolo 2, comma 1, del decreto medesimo, ente incaricato di dare attuazione, tenendo conto delle attività sviluppate nella fase di avvio e nei limiti delle risorse disponibili, al progetto *Human Technopole*, che si caratterizzi quale polo di ricerca innovativo, anche avvalendosi dei rapporti di collaborazione già instaurati, con la partecipazione del Ministero dell'istruzione, dell'università e della ricerca e del Ministero dell'economia e delle finanze, integrato con le università pubbliche dell'area metropolitana di Milano e con l'Istituto Italiano di Tecnologia, in coordinamento con le istituzioni e gli enti, anche territoriali, di riferimento, e in grado di realizzare le migliori sinergie con le reti nazionali e internazionali attive nei suoi ambiti di intervento;

VISTO il decreto del Presidente del Consiglio dei ministri 23 aprile 2015, con il quale al Sottosegretario di Stato alla Presidenza del Consiglio dei ministri, prof. Claudio De Vincenti, è stata delegata la firma di decreti, atti e provvedimenti di competenza del Presidente del Consiglio dei ministri;

SULLA PROPOSTA del Ministro dell'economia e delle finanze,

DECRETA

Articolo 1

(Approvazione del progetto esecutivo Human Technopole e individuazione delle risorse necessarie alla sua attuazione)

1. In attuazione di quanto stabilito dall'articolo 5, comma 2, del decreto legge 25 novembre 2015, n. 185, convertito, con modificazioni, dalla legge 22 gennaio 2016, n. 9, è approvato il progetto denominato *Human Technopole*, allegato al presente decreto.

2. E' autorizzato l'avvio del progetto di cui al comma 1, nel limite massimo delle risorse finanziarie previste dal citato articolo 5, comma 2, del decreto legge 25 novembre 2015, n. 185, che sono a tal fine assegnate all'Istituto Italiano di Tecnologia, che può avvalersi, altresì, delle proprie risorse umane e strumentali.

Articolo 2
(Attuazione del progetto)

1. L'Istituto Italiano di Tecnologia provvede, entro trenta giorni dalla data del presente decreto, all'avvio delle attività di realizzazione del progetto, fino all'operatività dell'ente di cui in premessa, da realizzare entro ventiquattro mesi dalla medesima data. A tal fine l'Istituto adotta le specifiche misure organizzative e le soluzioni gestionali dedicate, mediante una apposita Struttura di progetto, con contabilità separata. La struttura di progetto definisce gli aspetti logistici e organizzativi relativi alle operazioni di avvio della costituzione del polo di ricerca e provvede anche al reclutamento del personale necessario e all'acquisizione delle attrezzature scientifiche e tecniche necessarie nella fase inziale del progetto. Alla struttura di progetto è preposto il direttore di *Human Technopole*, che ne assume la piena responsabilità, scelto dall'Istituto con procedura selettiva internazionale tra persone di riconosciuta e comprovata esperienza e competenza, previo parere del Comitato di cui al comma 2.

2. E' istituito presso l'Istituto Italiano di Tecnologia, a valere sulle risorse destinate al progetto, un Comitato di coordinamento, per l'avvio della realizzazione del progetto *Human Technopole*, composto da due soggetti designati uno dal Ministero dell'economia e delle finanze, e uno dal Ministero dell'istruzione, dell'università e della ricerca; da tre scienziati di reputazione internazionale indicati di comune accordo dagli stessi Ministeri; dai rettori delle università pubbliche di Milano; dal presidente dell'Istituto superiore di sanità; dal presidente del Consiglio nazionale delle ricerche; dal presidente dell'Istituto Italiano di Tecnologia e dal direttore scientifico dell'Istituto Italiano di Tecnologia. Il Comitato si riunisce la prima volta su convocazione del Sottosegretario alla Presidenza del Consiglio dei ministri con delega per le valutazioni strategiche delle politiche pubbliche inerenti anche ai temi della ricerca scientifica e tecnologica, che presiede la riunione medesima fino all'elezione del Presidente. Nella stessa riunione, il Comitato elegge il Presidente tra i propri componenti e definisce le regole per il proprio funzionamento, anche in relazione all'esercizio delle diverse funzioni di competenza. Con decreto del Ministro dell'economia e delle finanze di concerto con il Ministro dell'istruzione, dell'università e della ricerca è stabilito l'ammontare massimo delle spese di funzionamento del Comitato ivi incluse quelle relative ai compensi e rimborsi spese da corrispondere ai membri del Comitato nel rispetto della normativa legislativa e regolamentare vigente in materia.

3. Il Comitato:

   a) verifica periodicamente, anche attraverso la misurazione di indicatori di performance collegati al raggiungimento degli obiettivi espressamente individuati nel progetto, l'effettiva coerenza tra il progetto *Human Technopole* e le attività svolte dalla struttura di progetto di cui al comma 1, con particolare riguardo all'efficienza, alla trasparenza e all'efficacia delle attività di gestione del progetto;

b) assicura il raccordo delle attività relative al progetto con le competenze di più elevata qualificazione, in ambito pubblico e privato, nell'area metropolitana di Milano, in Italia e su scala internazionale;

c) esprime pareri e formula indicazioni sugli interventi operativi necessari alla progressiva attuazione del progetto, predisposti o adottati dalla Struttura di progetto dell'Istituto Italiano di Tecnologia, con particolare riguardo all'allocazione delle risorse finanziarie e umane disponibili tra i centri di ricerca di cui all'allegato progetto *Human Technopole* e alla scelta dei rispettivi responsabili, alla definizione dei criteri di selezione del restante personale sulla base degli standard internazionali riconosciuti, in modo da assicurare l'apertura internazionale, la contendibilità degli incarichi, la trasparenza delle scelte e l'alta professionalità degli scienziati chiamati a ricoprire posizioni direttive e del personale di ricerca, nonché alle attività relative alla realizzazione delle grandi infrastrutture previste nel progetto *Human Technopole.*

d) rendiconta periodicamente del proprio operato, con apposita relazione semestrale, al Ministero dell'istruzione, dell'università e della ricerca e al Ministero dell'economia e delle finanze.

4. Il Presidente del Comitato e il direttore della Struttura di progetto di cui all'articolo 2 curano il raccordo operativo tra il progetto *Human Technopole,* la società Arexpo e gli altri enti pubblici e privati al fine di stabilire sinergie e collaborazioni scientifiche e tecnologiche.

Il presente decreto sarà trasmesso ai competenti organi per il controllo.

Roma, li 16 SET. 2016

p. IL PRESIDENTE DEL CONSIGLIO DEI MINISTRI
IL SOTTOSEGRETARIO DI STATO
(prof. Claudio De Vincenti)

IL MINISTRO DELL'ECONOMIA E DELLE FINANZE

CORTE DEI CONTI
UFFICIO CONTROLLO ATTI P.C.M.
MINISTERI GIUSTIZIA E AFFARI ESTERI
Reg.ne - Prev. n. 2749
17 OTT 2016

IL MAGISTRATO

PRESIDENZA DEL CONSIGLIO DEI MINISTRI
SEGRETARIATO GENERALE
UFFICIO DEL BILANCIO E PER IL RISCONTRO
DI REGOLARITA' AMMINISTRATIVO-CONTABILE
VISTO E ANNOTATO AL N. 1943/2016
Roma, 23.9.2016
IL REVISORE
IL DIRIGENTE

Ministero
dell'Economia e delle Finanze
UFFICIO DEL COORDINAMENTO LEGISLATIVO
Prot. 1571

Roma, 15/9/2016

Alla Presidenza del Consiglio dei Ministri
- Segretariato Generale

e, per conoscenza:

Al Gabinetto del Ministro

Al Dipartimento del Tesoro

Al Dipartimento della Ragioneria generale dello Stato

LORO SEDI

OGGETTO: Schema di decreto del Presidente del Consiglio dei Ministri, su proposta del Ministro dell'economia e delle finanze, in attuazione dell'articolo 5, comma 2, del decreto-legge 25 novembre 2015, n. 185, convertito, con modificazioni, dalla legge 22 gennaio 2016, n. 9. Istituto Italiano di Tecnologia. Approvazione del progetto esecutivo *Human Technopole.*

Si trasmette l'originale del provvedimento in oggetto indicato, debitamente bollinato dal Dipartimento della Ragioneria generale dello Stato e controfirmato dal Ministro dell'economia e delle finanze, con preghiera di sottoporlo, ove nulla osti, alla firma del Sottosegretario di Stato delegato.

Il Capo dell'Ufficio *ad interim*

Human Technopole ITALY 2040

Index of the Masterplan

PART 1 EXECUTIVE SUMMARY OF THE PROJECT

PART 2 TECHNICAL ANNEX

PART 3 FINANCIAL PLAN

Main Contributors to the Human Technopole Masterplan:

The Global masterplan of the Human Technopole was elaborated by a Committee constituted by IIT (Roberto Cingolani), Università Statale di Milano (Gianluca Vago), Università di Milano - Bicocca (Cristina Messa) and Politecnico di Milano (Giovanni Azzone).

The scientific masterplan of the Human Technopole was concieved and written by Roberto Cingolani, Pier Giuseppe Pelicci (Center C1, Facilities F1, F2, F4), Stefano Gustincich (Center C2, Facilities F1, F2, F4), Roberto Viola (Center C3, Facility F1), Mario Rasetti (Center C4), Andrea Cavalli (Center C5, Facility F3), Fabio Pammolli and Piercesare Secchi (Center C6), Guglielmo Lanzani (Center C7).

We acknowledge the contributions of colleagues (in alphabetic order) from many Institutions:

- **Onco Genomics Center (C1) and Facility F1 and F4**
  IIT Center for Genomic Sciences-Milano, IEO-Milano, INGM-Milano, INT-Milano, Politecnico di Milano, Humanitas - Milano, Università Statale – Milano, Universita' San Raffaele, (S. Abrignani, B. Amati , S. Ceri, P. Corradini, A. Mantovani, S. Minucci, G. Natoli, R. Orecchia, G. Tonon)
- **Neuro Genomics Center (C2) and Facility F1, F2 and F4**
  IIT Department for Neuroscience and Brain Technology - Genova, Università Statale – Milano, Università Bicocca – Milano, Besta – Milano, Humanitas – Milano Ospedale San Raffaele – Milano, Istituto M. Negri – Milano (M. Di Luca, C. Ferrarese, G. Forloni, G. Martino, M. Matteoli, F. Tagliavini)
- **Agri-Food and Nutrition Genomics Center (C3) and Facility F1**
  Università Statale - Milano, IEO – Milano, Humanitas – Milano, CREA – Roma, PTP – Lodi Università di Firenze, FEM – Trento (G. Carenzo, T. Cattaneo, L. Cattivelli, D. Cavalieri, M. Donatelli, A. Gentile, A. Mantovani, F. Mattivi, M. Pisante, M. Rescigno, G. Sacchi, A. Stella, C. Tonelli, R. Velasco)
- **Data Science Center (C4)**
  Institute Scientific Interexchange (ISI) - Torino, Università Statale – Milano, Università Bicocca – Milano, Politecnico – Milano, Università di Pavia, Università di Bologna (M. Canfora, S. Ceri, P. Contucci, E. D'Angelo, C. La Porta, A. Marzuoli, G. Mauri, G. Pasi, R. Ricca, P. Secchi, S. Zapperi)
- **Computational Life Sciences Center (C5) and Facility F3**
  IIT Compunet-Genova, Università Bicocca – Milano, Politecnico di Milano, CINECA – Bologna (S. Bassini, C. Cavazzoni, M. De Vivo, C. Di Valentin, G. Mauri, G. Pacchioni, W. Rocchia, D. Sciuto)
- **Center for Analysis, Decisions, and Society (C6)**
  Politecnico di Milano (M. Arnaboldi, M. Calderini, D. Sciuto, L. Tanca)
- **Center for Smart Materials and Devices (C7)**
  Istituto Italiano di Tecnologia, IIT- Center for Nano Science and Technology - Milano, Università Bicocca – Milano (I. Bayer, P. Branduardi, D. Fragkouli, M. Labra, P. P. Pompa, R. Simonutti)

The following **Acronyms** will be used in the document:

Besta: Istituto Neurologico Besta
CREA: Consiglio Ricerca in Agricoltura e l'Analisi dell'Economia Agraria
FEM: Fondazione Edmund Mach
Humanitas: Istituto Humanitas
IEO: Istituto Europeo di Oncologia
IIT: Istituto Italiano di Tecnologia
INGM: Istituto Nazionale Genetica Molecolare
INT: Istituto Nazionale Tumori
ISI: Institute for Scientific Interexchange (ISI-Foundation)
M. Negri: Istituto Farmacologico Mario Negri
OSR: Ospedale San Raffaele
PTP: Parco Tecnologico Padano

AD: Alzheimer's Disease
ALS: Amyotrophic Lateral Sclerosis
HT: Human Technopole
NHS: National Healthcare System
PD: Parkinson's Disease
PI: Principal Investigator
RL: Research Line

OGC: Onco Genomics Center (C1)
NGC: Neuro Genomics Center (C2)
AFNGC: Agri-Food and Nutrition Genomics Center (C3)
DSC: Data Science Center (C4)
CLSC: Computational Life Sciences Center (C5)
CADS: Center for Analysis, Decisions, and Society (C6)
CSMD: Center for Smart Materials and Devices (C7)

F1: Central Genomics Facility
F2: Imaging Facility
F3: Data Storage and High-Performance Computing Facility
F4: Common Shared Services Facility

Human Technopole is a registered logo.

PART 1 - EXECUTIVE SUMMARY OF THE PROJECT

1. INTRODUCTION

Italy has one of the world's highest life expectancies (82.3 years in 2012 according to OECD Health Statistics 2014) and high nutrition standards. What accounts for this remarkable life expectancy and how it can be improved is still an open question. Health, aging, and quality of life are affected in a complex way by a combination of *intrinsic* factors, primarily related to each individual's genetics, and *extrinsic* factors, such as nutrition, lifestyle, and environment. A comprehensive approach to health and aging (hereafter referred to as ***Human Technologies***) does not yet exist, in part because this would require that cutting-edge technologies be integrated with high-profile basic and translational science in the critical areas of Medicine, Data Science, Nanotechnologies, and Nutrition.

Italy's challenge (namely *Italy 2040*) is to become a world leader in *Human Technologies* by embarking on an intensive cross-disciplinary project (Fig. 1) to synergistically develop fundamental and clinical genomics, nutrition, innovative algorithms for data analysis, multiscale methods in computational life sciences, and advanced technologies for food and diagnostics. To achieve this, we propose to create **a national cross-disciplinary research infrastructure named 'Human Technopole' (HT)** in Milan's EXPO site. **The Human Technopole's mission will be to develop personalized approaches, both medical and nutritional, focusing on cancer and neurodegenerative diseases. It will achieve this mission using genomics, the analysis of increasingly large data sets, and new diagnostics techniques.**

HT will pursue different major lines, including:

- the construction of a knowledge-based environment for the development of Genomic Sciences and the rapid and cost-effective translation of discoveries into patient benefits and industrial applications;
- the development and integration of state-of-the-art genomic technologies with advanced basic and translational research, big-data analyses and cutting-edge clinical care to tackle some of the most relevant threats to human health in ageing individuals, namely cancer and neurodegenerative diseases;
- developing healthier and safer food through integrative genomics and systems biology and by adopting new sustainable technologies for food production, conservation, and storage;
- implementing powerful methods for artificial intelligence and statistics to extract knowledge from data in order to facilitate an efficient Precision Medicine program;
- linking systems biology and network pharmacology using big data analysis and new predictive algorithms, and developing innovative multiscale approaches to computational biology, drug discovery, and health;
- processing the massive amounts of socioeconomic data available using high-performance computing and storage facilities so that innovative analytical solutions can be developed for public decision makers;
- developing new-concept fast, cheap, disposable devices for sensing, diagnostics, and high-throughput screening of biological samples (both patient and food samples);

The HT program's expected results will be:

- The creation of a large-scale international research infrastructure, located at Milan's EXPO site, to give continuity to the original idea of *feeding the planet* and to increase Italy's leadership in fields dealing with quality of life. The facility will involve about 1500 people and will include:
    - 30000 sqm of cross-disciplinary laboratories including 7 Centers and 4 Facilities;

    - Massive high-throughput genomic screening/sequencing;
    - Imaging dedicated to structural biology and proteomics;
    - Data storage and a high-performance computing center.
- The first large-scale national screening (carried out in a countrywide network of research hospitals) of:
    - cancer patients for treatment stratification;
    - patients with neurodegenerative diseases for genomic-based stratification;
    - healthy people for disease risk-assessment and prevention;
    - cancer patients for identification of new markers of sensitivity or resistance to innovative therapies.
- The recruitment of more than 1000 scientists (including about 100 principal investigators, researchers, technologists, postdocs etc.) exclusively *by international calls*, with a strong reverse brain-drain effect;
- The production of new scientific knowledge and new technologies for diagnostics and sustainability in the fields of health, nutrition, data analysis, computational life sciences, and nanotechnology;
- The development of predictive models that can be applied to large ensembles of data in a variety of domains, from health to other sectors of great public interest;
- The development of statistical, mathematical, and economic models that will support a new generation of data-aware public policies;
- The strengthening of public-private partnerships and industrial collaborations in fields of strong social impact and great economic relevance for the country, namely: public health, food and nutrition, data analytics, and diagnostics;
- A large-scale educational program (at PhD level) in the Human Technopole's research fields;
- A national outreach and dissemination program on science, health, and technology in collaboration with several national and international players .

The Human Technopole in Milan will trigger the development of a national precision medicine program in collaboration with the Istituto Superiore della Sanità (Ministry of Health) and with Regional Governments across the country.

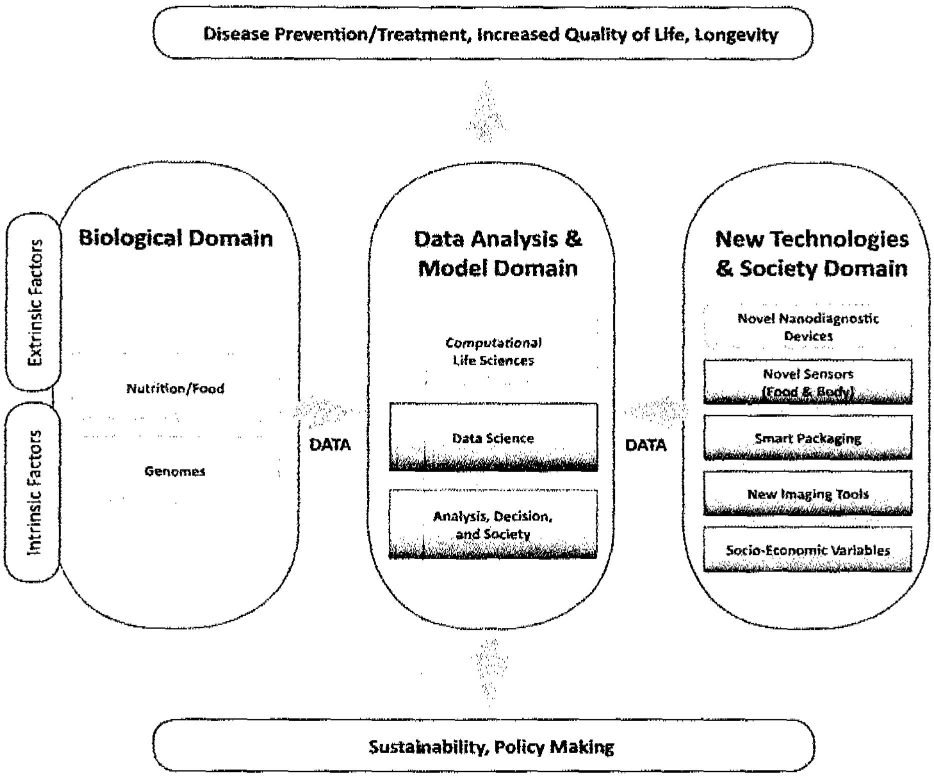

Disease Prevention/Treatment, Increased Quality of Life, Longevity
Extrinsic Factors
Intrinsic Factors
Biological Domain
Nutrition/Food
Genomes
DATA
Data Analysis & Model Domain
Computational Life Sciences
Data Science
Analysis, Decision, and Society
DATA
New Technologies & Society Domain
Novel Nanodiagnostic Devices
Novel Sensors (Food & Body)
Smart Packaging
New Imaging Tools
Socio-Economic Variables
Sustainability, Policy Making

Figure 1. Schematic representation of the project.

2. STATE OF THE ART

Cancer and neurodegeneration are the most significant causes of death and detrimental aging in advanced societies. The WHO has calculated that, in 2030, 13 million people worldwide will die of cancer. In 2014, 366000 new cancer cases (1000 per day) were registered in Italy alone. In 2008, 2.45 million people were diagnosed with cancer and 1.23 million died from cancer in the 27 countries of the European Union (EU). In the EU, cancer costs reached €126 billion in 2009, with healthcare accounting for € 51 billion (40% of all cancer costs).

In addition, 35 million people worldwide suffer from a neurodegenerative disease. That number is likely to rise to 100 million in 2050. Disease progression is slow, resulting in a massive economic impact on society due to the costs of assistance and the disruptive consequences for families. The total cost for dementia care in Europe is €177 billion/year, with €80.8 billion spent on direct institutional care and €96.6 billion spent on informal care. In Italy, these costs are €4.8 and €15 billion, respectively (2008 data). Italy has one million patients with Alzheimer's disease (AD) and 200000 patients with Parkinson's disease (PD). AD, PD, Huntington's disease (HD), and amyotrophic lateral sclerosis (ALS) are all incurable. AD was recently recognized by the European Health Committee as a health policy priority.

Nutrition affects life expectancy and the quality of aging in advanced societies, with diet being an important contributor to both the development and the prevention of diseases. Aging and the rate of aging-associated diseases (cancer, diabetes, atherosclerosis, cardiomyopathies, autoimmune diseases, neurodegenerative diseases, respiratory diseases, and kidney diseases) are controlled by conserved genetic pathways, which are regulated by food intake and lifestyle. These factors influence several critical cellular functions by inducing changes in metabolism and gene expression. These pathways can be modulated pharmacologically or through the diet to increase longevity and reduce the incidence and severity of aging-associated diseases. The process should be monitored by specific (digital) genomic and epidemiological methodologies.

Understanding and exploiting the cross-correlations between nutrition, genetics, aging, and life expectancy would improve both healthcare and food technology, resulting in ***preventive nutrition*** and ***personalized medicine*** for citizens. This would positively impact the quality of life for future generations as well as the performance of the public health system. It would create new opportunities:

- *For science:* to take novel approaches in developing new technologies and in studying fundamental biological questions as well as new modelling and big data analytics methods;
- *For patients:* to improve risk assessment, diagnosis, prevention, and healthcare;
- *For the economy:* to pursue industrial development, particularly in the fields of food technology, models and software, diagnostics and therapies;
- *For public health:* to improve outcomes, to increase public expenditure efficiency, and to improve the population's quality of life and life expectancy;
- *For policy and decision makers:* to provide public decision makers with the analytical methods and predictive algorithms to better analyze, check, and predict complex socioeconomic scenarios, allowing them to make more informed decisions.

Worldwide situation: The impact of genomics in healthcare has recently attracted attention in both the public and private sectors.

1. In 2004, the **Broad Institute** in Boston (US) was jointly launched by the Broad Foundation, MIT, and Harvard to study genomics and systems biology in correlation with various diseases (primarily cancer);
2. In late 2012, England launched the 100000 **Genomes Project**. Genomics England, a company wholly owned and funded by the Department of Health, was set up to deliver this flagship project, which will sequence 100000 whole genomes from patients by 2017;
3. The **National Institute of Health (NIH, US),** working with the Cancer Genome Initiative and the Wellcome Trust Sanger Institute's Cancer Genome Project (United Kingdom), has already sequenced thousands of cancer cell genomes;
4. In the US, the **Alzheimer's Disease Sequencing Project** has begun releasing data from the peripheral tissues of hundreds of patients;
5. **deCODE genetics** has gathered genotypic and medical data from more than half of Iceland's adult population to identify risk factors for several human diseases;
6. In 2015, Barack Obama launched the **Precision Medicine Initiative** to recruit a research cohort of one million or more volunteers in order to foster genomic data and scale up efforts to identify genomic drivers in cancer that can be used to develop new drugs;
7. **ELIXIR Centers Europe** is leading life science organizations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. It is a pan-European research infrastructure for biological information;
8. **Google Genomics** has been funded to provide IT infrastructures to store, process, explore, and share genomic data using Google's cloud infrastructure;
9. **Global Alliance for Genomics & Health (Precision Medicine)** has grouped research centers, hospitals, and pharmaceutical companies from 34 countries to facilitate and catalyze the sharing of genomic and clinical data in an effective, responsible, and interoperable manner;
10. In 2013, the Obama Administration launched an extended **Big Data R&D Initiative**, committing more than US$200 million in new funding through six agencies and departments to improve "our ability to extract knowledge and insights from large and complex collections of digital data";
11. **Oxford Parkinson's Disease Detection Dataset** is a Parkinson's Telemonitoring huge dataset run by artificial intelligence machine-learning voice recognition methods. It collects information on PD patients in real time and is the largest European repository of patient data produced using various brain imaging techniques;

12. In China, the **Bejing Genome Institute** has launched the 1000 Plant Genomes Initiative, inviting the most advanced institutions worldwide to collaborate in decoding all the major crops. It is exploring the potential of Genome Wide Association analysis to implement assisted breeding and breeding by design using cultivar resequencing and gene-to-trait association;
13. **Genome Canada** invests between $50 million and $100 million across public and private sectors each year to find new uses for genomics. It invests in large-scale science and technology to fuel innovation and to translate discoveries into applications, new technologies, societal impacts and solutions across key sectors of national importance (health, agriculture, forestry, fisheries & aquaculture, energy, mining, and the environment);
14. The Smithsonian Institute's **Global Genome Initiative (GGI)** is a collaborative effort to create a solid foundation for genomic research through a global network of biorepositories and research organizations. The GGI will preserve and study genomic diversity and increase access to genomic information from the key branches of the Tree of Life, expanding its contribution to the preservation and knowledge of life on our planet;
15. *In Europe, the* **JPI-ENPADASI** *initiative in nutrigenomics and food science is a partnership to* connect the major European Centers focused on nutrition, food, and health by establishing good practices and methods for data sharing and data integration at the EU level;
16. Worldwide, several Centers have started to combine methodological research, data-driven research, and domain-specific research in order to address important socioeconomic challenges in areas such as energy, finance, health delivery, epidemiology, economics, and management science. The Centers are: the **MIT Institute for Data, Systems, and Society**, the **Network Science Institute** at Northeastern University, the **Stern Center for Business Analytics** at New York University, the **Big Data Group** at Cambridge University, the **Alan Turing Center** in London, the **iLab at Carnegie Mellon University**, the **Center for Data Science and Public Policy** at University of Chicago, and the **Berlin Big Data Center.**

3. GENERAL DESCRIPTION

Within the international context outlined above, the Human Technopole seeks to:

i) create a large-scale scientific infrastructure and a collaborative public/private ecosystem at the EXPO site, adopting the highest quality standards to attract the best talent;

ii) develop an integrated approach to genomics, nutrition, diagnostics, and data analysis in order to prevent and treat those diseases which most greatly impact the healthcare system (namely, cancer and neurodegeneration);

iii) produce every year (in the early years) about 2000 genomic screenings of healthy individuals for disease risk-assessment and prevention, about 2000 genomic screenings of cancer patients for treatment stratification, about 1000 cancer patients for identification of new markers of sensitivity or resistance to innovative therapies, about 1000 genomic screenings of patients with neurodegenerative diseases, and about 1000 screenings of biomarkers for diagnostic purposes. After the start-up phase, these numbers will increase substantially thanks to the progressive involvement of research hospitals and universities in different regions (with a competitive target of 30000 genomes per year);

iv) develop new technologies to improve the fields of medicine and nutrition.

The Human Technopole's long-term goal is to make Italy a leading player in the use and development of Precision Medicine. Its novel and comprehensive approach will synergize medical science with the impact of Italian food and nutrition on human health and healthy aging.

Italy 2040 is naturally a long-term vision and its progressive implementation will have a lasting impact on our society for the next 25 years and beyond. To fully implement this vision, strong synergy is required

between research institutions working in different fields. HT will thus create a network of collaborations with public and private research and clinical institutions in Italy and abroad. It will establish **joint research programs, joint laboratories**, and **outstations** in research hospitals and research entities orbiting around the Human Technopole program (Outstations are HT laboratories established outside the Expo Headquarter in relevant scientific/clinical hosting institutions which are operated with HT resources by HT scientists, based on a collaboration agreement with the hosting institution).
This is particularly important in view of the unprecedentedly broad effort required to create a unified national strategy by merging clinical studies, fundamental genomics, nanotechnology, data science, high-performance computing, and social analyses.

The Human Technopole will be a new legal entity to be incorporated at the beginning of 2017. The project will have a start-up phase of two years, with the target of:

1) Initiating the set-up of the main infrastructure;
2) Starting the collaborations network;
3) Starting the recruitment of senior scientific staff, exclusively by international calls;
4) Organizing the core of the HT administration offices.

During the start-up phase, IIT will be in charge of implementing the above items, supervised by a high-level Governmental Advisory Board. This masterplan presents the Human Technopole's scientific and organizational development in the first 7 years. Importantly, the masterplan forecasts are indicative and will need continuous updating as recruitment evolves, depending on the detailed scientific plans of the Principal Investigators hired.

The Start-Up Phase
The start-up phase's main targets are detailed below:

1. **To initiate the main infrastructure set-up, in part by renovating existing buildings at the EXPO site;**

The Human Technopole will begin with 7 main Centers and 3 large-scale Facilities (see Fig. 2) located at the EXPO site in Milan. **Centers** will have their primary infrastructure at the EXPO site, but may also have **Outstations** located outside the EXPO site according to specific interinstitutional agreements (e.g. in Research Hospitals).
Each Center may comprise one or more laboratories and will pursue several well-defined Research Lines (RL), as described in Part 2. Research Lines will be pursued autonomously by newly hired researchers in collaboration with Universities or other research institutions.
The 7 Centers and 3 Facilities located at the EXPO site will be (see Fig. 2):

C1: Onco Genomics Center;
C2: Neuro Genomics Center;
C3: Agri Food and Nutrition Genomics Center;
C4: Data Science Center;
C5: Computational Life Sciences Center;
C6: Center for Analysis, Decisions, and Society;
C7: Center for Smart Materials and Devices.

These will be supported by 3 main Scientific Facilities:

F1: Central Genomics Facility;
F2: Imaging Facility;
F3: Data Storage and High-Performance Computing Facility;

Laboratories and services of common interest will be grouped in a fourth Facility (F4: Common Shared Services Facility).

2. **To begin building the collaborations network, primarily with clinical partners;**

During the start-up phase, HT will rely on a collaborations network involving several institutions with complementary expertise (mostly in the Milan area), namely:

- Istituto Italiano di Tecnologia, Milan's three public universities (Politecnico, Università Statale, Università Bicocca);
- Istituto di Ricerche Farmacologiche M. Negri, Istituto Nazionale Genetica Molecolare, CREA (Consiglio Ricerca in Agricoltura ed Analisi Economia Agraria);
- A network of high-level Research Hospitals including: Istituto Europeo di Oncologia (IEO), Istituto Nazionale Tumori, Humanitas, Istituto Neurologico C. Besta, Ospedale Maggiore Policlinico, Ospedale Universitario San Gerardo - Milano Bicocca, Ospedale San Raffaele;
- The national supercomputer facility CINECA, the Institute for Scientific Interexchange (ISI-Foundation) - Torino; Fondazione Edmund Mach (FEM) – Trento;

For such a national long-term enterprise, these partners have been strategically chosen for their excellent skills in complementary fields. An additional factor is that they can supply high-quality PhD students (the Universities) and clinical data (the Research Hospitals). As the start-up phase progresses, additional Italian institutions will become involved via periodic calls and/or interinstitutional agreements for proposals to contribute to the Human Technopole's research plan. These instituations may include CNR, other Universities, research hospitals and companies country wide. Figure 2 exemplifies the collaborations network between the different Institutions participating in the Human Technopole's start-up phase.

Most Centers will have Outstations where a portion of the research and development will be conducted. The Onco Genomics Center and the Neuro Genomics Center will form a network of Outstations involving some of the most important Research Hospitals in both the Milan area and nationwide via the Istituto Superiore di Sanità. This network will ensure access to medical knowledge, biological samples, and clinical data as well as allowing proper data storage in a high-level unified national database and accelerating clinical trials.

Fundamental collaborations will be developed with Milan's three main public Universities. These collaborations will be based on joint activities and/or laboratories hosted within the Centers as well as outside the Centers. The core collaborations with the Statale di Milano will primarily deal with the Onco Genomics Center, the Neuro Genomics Center, and the Agri Food and Nutrition Genomics Center. The Center for Computational Life Sciences, the Center for Neuro Genomics, and the Center for Smart Materials and Devices will develop joint activities with Università Bicocca. The Center for Analysis, Decisions, and Society and the Center for Smart Materials and Devices will develop joint activities with Politecnico di Milano. The Data Storage and High-Performance Computing Facility will be located at the national supercomputer center in Bologna (CINECA), and will serve HT and the entire network of collaborating institutions.

Importantly, in addition to this team of core partners, other national and international public research entities have officially expressed interest in developing collaborations with HT. These includethe European Molecular Biology Laboratory (EMBL), INAIL (Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro; the Italian National Work-Accident Insurance Company), the Istituto Superiore di Sanità, the ACC national network of IRCSS, and several Regional Governments (Lombardia, Emilia Romagna, Liguria, Umbria). Contacts with other Regions are already on going (Puglia, Lazio, Piemonte).

Finally, several private organizations have expressed interest in the Human Technopole. These organizations include: Assolombarda, several agro-food companies, IBM Research, ST Microelectronics, and several important charitable foundations such as Fondazione Cariplo, Compagnia San Paolo, Fondazione Don Gnocchi, Fondazione Feltrinelli, Fondazione Golinelli, Fondazione Veronesi, and Fondazione Altagamma.

Collaborations and joint initiatives with these entities will be crucial for creating an international public-private ecosystem at the EXPO site and for driving the Human Technopole's future development. They all will be discussed and developed during the start-up phase. A list of Letters of interest is enclosed in the Appendix 1 of this Part 1.

3. **To begin recruiting senior scientific staff (including Chief Scientist, Center Directors, and an initial core of Principal Investigators and Staff Researchers) exclusively by international calls;**

Recruitment at HT will be organized according to international standards and conducted exclusively via international calls, as described in Section 6. The start-up phase's first recruitment will be the Chief Scientist, an internationally acknowledged scientist to take charge of building the Human Technopole and implementing the Scientific Strategy. He/she should be recruited by an *ad hoc* search committee, through a confidential Expression of Interest procedure. The start-up phase's second recruitment action will be for Center Directors and Facility Directors. These hirings will be conducted through international calls (in a similar way to the IIT tenure track procedure). *Ad hoc* International Search Committees will be convened in different scientific areas to create shortlists and interview selected candidates. Once the Chief Scientist and the Directors of the Centers and Facilities are appointed, the scientific staff will be recruited. Tenure track researchers, staff researchers, technologists, and post docs will be recruited exclusively by international calls. PhD students will be recruited according to standard rules.

4. **To organize the core of the HT administration offices;**

During the start-up phase, an administrative core will be set up to ensure the HT initiative is correctly launched in accordance with the schedule for the initial fundamental activities, namely:

a. launch the calls for and recruitment of scientists (chief scientist, center coordinators, tenure track scientists, staff scientists, postdocs, technicians);
b. tenders and purchases;
c. infrastructure implementation;
d. website;
e. finances;
f. Institutional agreements/contracts/collaborations.

In the early stage, while the HT administration offices are still being put together, the above activities will be supported by IIT administrative staff from the following offices, temporarily located at HT:

- Human Resources and Organization and Research Organization Office (point a, f);
- Procurement and Purchasing Office (point b);
- Financial Planning and Control (all points);
- Technical Services and Facilities, Health and Safety, and Information and Communication Technology (point c);
- Communication and External Relations (point d);
- Administrative Management, Management Control Office (point e);
- Technology Transfer Office (point f).

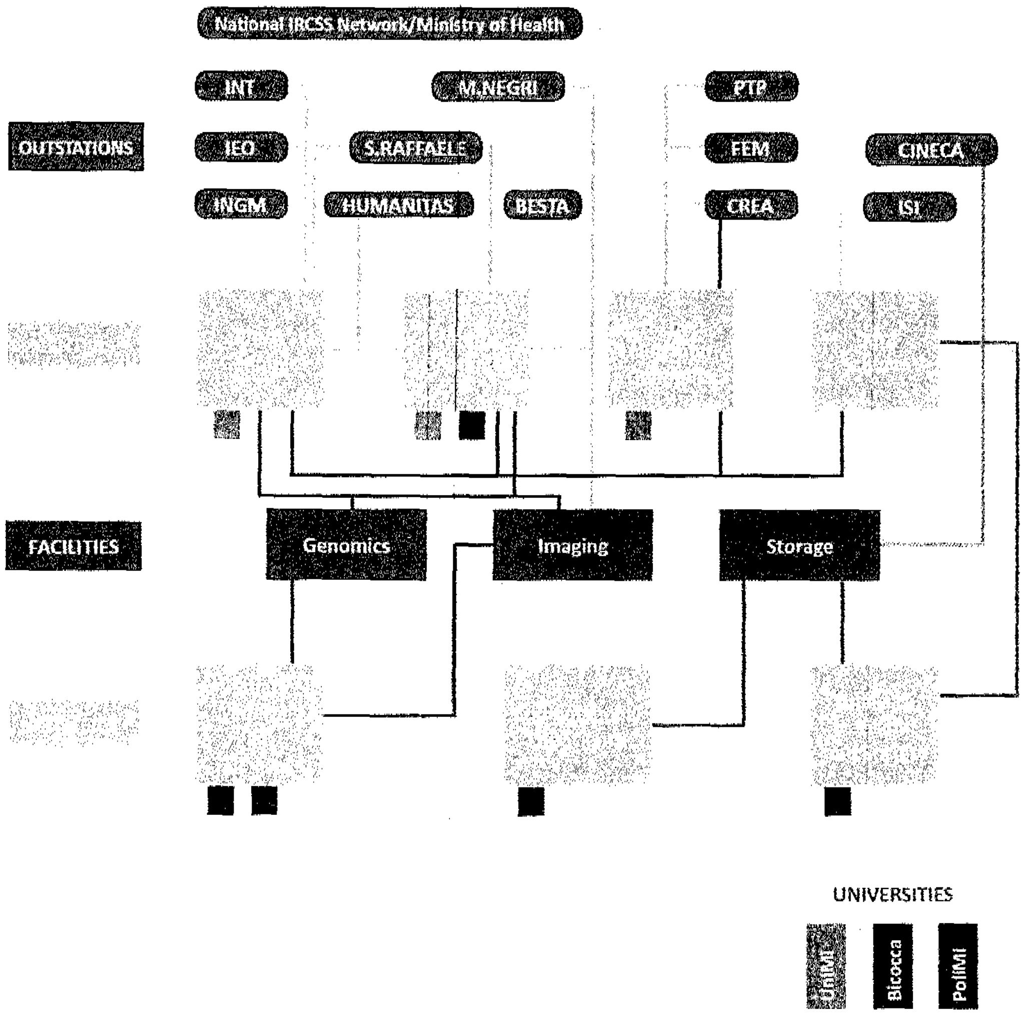


Figure 2: Structure of the Human Technopole at the start-up. The Yellow boxes represent the Human Technopole's 7 Centers. Centers will conduct joint research activities with Milan's public Universities (Red bar: PoliMi, Light Blue bar: UniMi, Dark Blue bar: Bicocca) and with Institutions, where joint Outstations will be established outside the EXPO site (Purple boxes). All Centers will benefit from the shared Facilities for Genomics, Imaging, and Storage/HPC (Green boxes).

The first assessment of the Human Technopole and the Centers is planned for the end of the start-up phase.

After the start-up phase, HT could develop other activities dedicated to young researchers and to scientific outreach and dissemination, namely:

1) **Seed Projects**

In addition to in-house research, the Human Technopole could open periodic **public calls for projects** dedicated to researchers around the country, bringing in new ideas and methods that are of interest for the Human Technopole program. These projects are intended to last for 2 years with funding mechanisms similar to those of junior research grants.

2) **Outreach and Dissemination**

A complete outreach and dissemination program associated with the Human Technopole's science and technology activities . This program should include the following initiatives:

- Digital and media dissemination of science (editorial agreements with national public TV network and with major newspapers);

- High school dissemination programs;
- Update courses for high school teachers;
- Partnerships with major national dissemination events to promote the Human Technopole's message (e.g. Festival della Scienza, Bergamo Scienza, and Repubblica delle Idee).

The Outreach and Dissemination program should be conducted in collaboration with major national editorial partners such as Fondazione Feltrinelli, Fondazione Golinelli, and Fondazione Altagamma.
Charities and patients' associations dedicated to cancer and neurodegenerative diseases will also be important partners in the dissemination and outreach campaign.

3) **HT PhD School**

A cross-disciplinary PhD course needs to be established, with courses and experimental projects focusing on the core scientific disciplines involved in the HT research program.

4. EXPECTED IMPACT, MILESTONES, AND PROJECT SCHEDULE

The basic ingredients of **Precision Medicine** and **Preventive Nutrition** are the big data analysis of the genetic profiles of large patient populations in correlation with the clinical, environmental, and lifestyle data of these populations. Together, they can reduce national health costs and improve quality of life. The Human Technopole will thus have a fundamental scientific focus as well as great technological potential in different contexts. Life scientists, clinicians, computer scientists, nanotech scientists, and data scientists will closely interact to achieve the shared goal of developing Precision Medicine approaches to cancer and neurodegenerative diseases. This will boost fundamental scientific knowledge and novel health-protecting strategies, including personalized prevention plans and treatments.
Data Science will allow us to merge Precision Medicine with data-based predictions, simulations, and scenario generations in order to provide decision makers and policy makers with new tools for developing rational strategies. We will develop novel algorithms to extract correlation patterns from large data sets. This will lead to more extensive information and, eventually, more profound knowledge of the mechanisms and processes of medicine and life science systems. By combining large-scale computational facilities, predictive modeling capabilities, and domain-specific competences and models, the HT will support significant innovation in analyzing socioeconomic systems.

For patients: the project proposes a brand-new modern approach to personalized therapy and prevention. It will provide a solid foundation for population screening and disease prevention. At the same time, it will optimize the decision-making process in therapy and decrease the current side-effects of several drugs and treatments for some of the greatest threats to human health.

For the economy: the project will provide Italy with a competitive advantage in reversing the brain drain, attracting industrial partners and investment (ICT, Food, Pharma and Biotech), and improving treatments in hospitals and clinics in terms of both clinical efficacy and cost containment. HT has the potential to create a favorable environment for start-ups: from software houses for analyzing genomic, clinical, and socioeconomic data to biotech companies producing tools, improved instruments, new drugs, and new diagnostics. Finally, the project will provide new ICT tools and predictive statistical methods for application in a variety of fields of great social importance (from health to public administration).

For decision makers: The Human Technopole's big data and information processing capabilities will provide a strengthened analytical background to support decision makers and policy makers in the following domains: *(i)* healthcare and welfare; *(ii)* science, technology, and industrial policy; *(iii)* analysis, management, prediction, and control of complex economic, social, and financial interactions; *(iv)* decisions and policies made by public and private institutions operating in complex domains.

The creation of the Human Technopole will follow a precise roadmap, based on the scientific plan described in Part 2. For the start-up phase (the first 2 years), the main objective will be to initiate the international recruiment of PIs and to build the main lab infrastructure, including laboratories, offices, and services (workshops, parking, warehouses etc.). The following Gantt (Fig. 3) displays the Human Technopole's tentative milestones for the first 4 years. Notably, this forecast may change depending on logistical issues, financial issues, and the recruitment schedule. Taking into account these caveats, we can safely assume that refurbishing the HT buildings will drive the time Gantt. The first heavy laboratories could be set up after approximately 24 months, with a steady state condition reached for the infrastructure after about 4 years. The growth in staff numbers (see 'headcount forecast' in section 7) is compatible with this logistical scenario. During the first 24 months, activities will be possible at some HT Outstations and, for equipment-light or computational research, at the EXPO site (shadowed lines in Fig. 3).

The following milestones are expected to be accomplished:

Year 2:

1. 50% of the infrastructure initiated or completed;
2. Outstations and collaboration network agreements completed;
3. Recruitment of directors completed;
4. Tender for laboratory instruments launched/completed.

Year 4:

1. Centers and Facilities completed;
2. Infrastructure completed;
3. Recruitment according to schedule (see section 6);
4. Fund raising activities started;
5. >5000 screening/year performed.

A more detailed Gantt with Milestones and Deliverables will be produced uring the first year together with the founding partners of the new legal entity "Human Technopole".

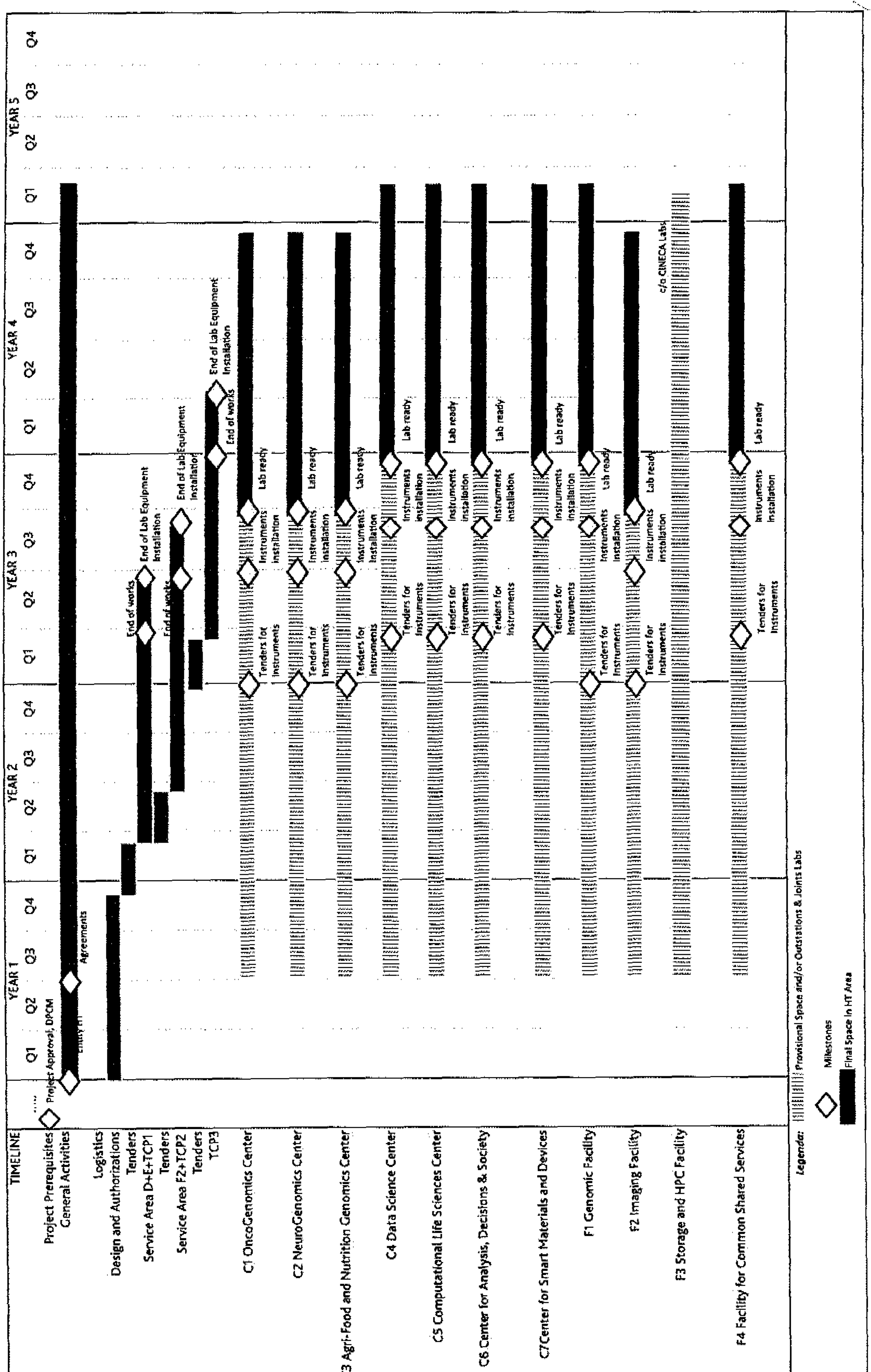


Figure 3: Preliminary Gantt for the Human Technopole.

5. INFRASTRUCTURE

The Human Technopole will consist of a facility **of approximately 30000 sqm,** including research laboratories, service labs, common areas, meeting and seminar rooms, and administration offices. This structure will host the 7 Centers, the 3 Facilities, common shared services (F4), and some of the Joint Laboratories with the Universities, as described above. The infrastructure's essential design must privilege functionality, safety, and energy saving. Important requisites are:

- About 3 MW power available;
- Continuity for at least 1 MW;
- Fast interconnection and ICT infrastructure;
- Antivibration platforms;
- EM-shielded areas.

Since the Human Technopole will attract national and international partners (universities, research centers, and companies), **additional logistics will be necessary to host teaching activities, Joint Laboratories and Outstations for external institutions, and Outreach Inititatives.** Finally, **space should be made available for start-ups and industrial laboratories.** At this stage, it is difficult to predict the needs and the timing of such collaborative initiatives. However, it is reasonable to start with a **buffer of available space in the range of 5000 sqm** to attract the first group of partner institutions. Further expansion can be postponed until after the start-up phase within the context of better defined and established project financing. **We envisage a final site comprising at least 35000 sqm of floorspace, where scientists, students, companies, and research teams will work together.**

A quick start-up for the Human Technopole can be achieved by refurbishing the preexisting infrastructures at the EXPO site.
Preliminary site inspections have revealed that several EXPO buildings located between the "Cardo" and the "Decumano" (the EXPO site's main streets) could be suitable for adaptation and refurbishment (Fig. 4).

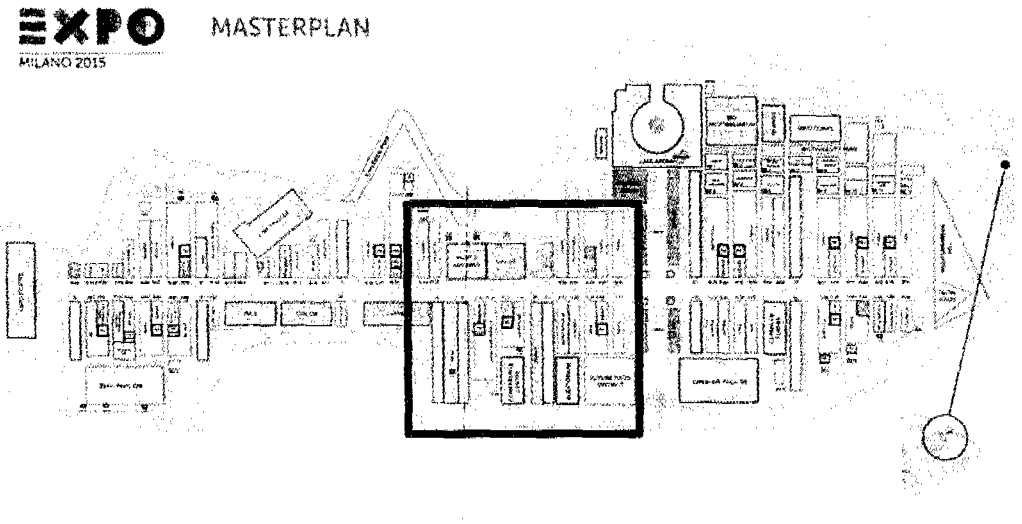


Figure 4: Areas, pavilions, and buildings which could be adapted for the Human Technopole.

Figure 5 provides a detailed map of the area highlighted in Figure 4. The three so-called "Service Area" buildings (long orange structures in Fig. 5) each comprise two levels (4000 sqm plus basement) for a total useful area of about 12000 sqm. These buildings could be adapted with relatively little effort in about 24 months for a quick Human Technopole start-up. The three TCP buildings each have an area of 2000 sqm and are approximately 12 meters tall. Multiple-level structures could thus be designed for a total of approximately 6000 sqm of floorspace for each building (3 levels, 4000 sqm for 2 levels), resulting in a maximum of 16000 sqm in about 36 months. Finally, new buildings could be constructed to produce more than 5000 sqm of floorspace.

We envisage laboratories on the ground floors (6000 sqm in the three TCP buildings, and 6000 sqm in the three Service Area buildings) and office spaces on the first floors of the 6 buildings. Assuming approximately 150 desks per floor plus common areas and services , we might consider 900 desks in the first 4 years. If a third level is built in each TCP, this would make additional office space available. Alternatively, additional office space could be provided by constructing a new building in the available area.

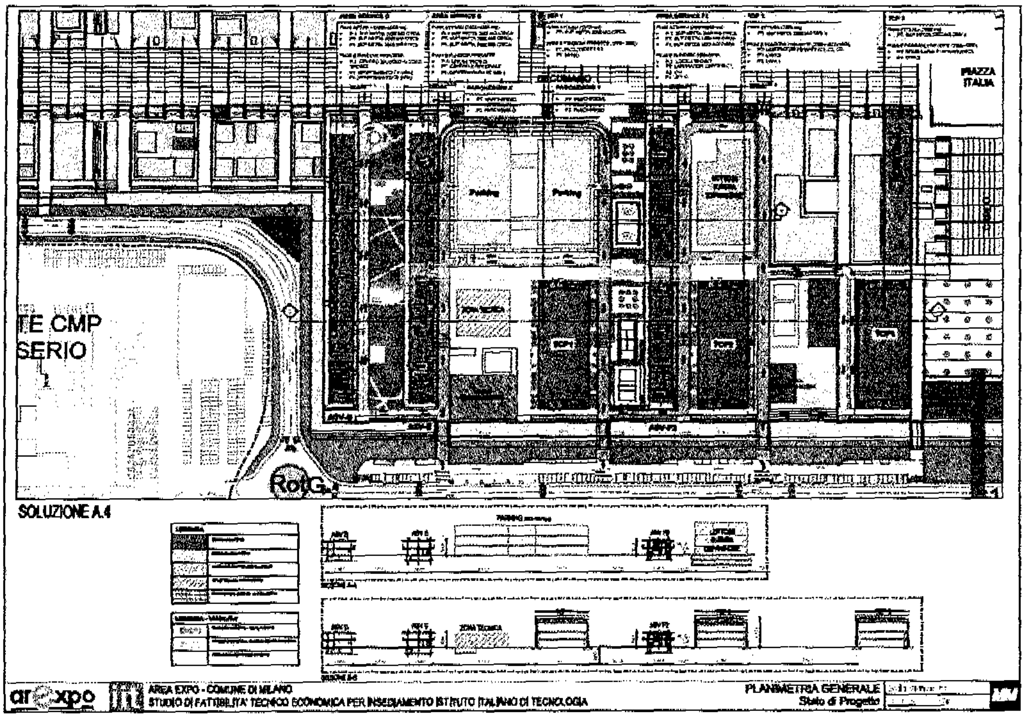

PIAZZA ITALIA
TE CMP
SERIO
RotG.
SOLUZIONE A.4
arexpo
AREA EXPO - COMUNE DI MILANO
STUDIO DI FATTIBILITA' TECNICO ECONOMICA PER INSEDIAMENTO ISTITUTO ITALIANO DI TECNOLOGIA
PLANIMETRIA GENERALE
Stato di Progetto

Figure 5: Areas, pavilions, and buildings which could be adapted for the Human Technopole.

Notably, while the available buildings are being refurbished in the early phase, a first allocation of offices and light laboratories (for about 250 people) could be provisionally located in the "Cascina Triulza" building (about 1250 sqm) and in the "Intesa-San Paolo Pavilion" (about 1900 sqm), which would be donated to HT by Banca Intesa-San Paolo. In addition, a portion of the scientific activity could begin at the Outstations and Joint Laboratories. This is consistent with the expected headcount growth (see section 6).

Before concluding this section, we would like to point out that the final decision about the HT location will be made at a later stage, in coordination with Arexpo Company, the owner of the areas and the buildings, within the masterplan of the entire Expo Area. In this frame Arexpo will also help to identify new areas (possibly near the HT buildings), for the location of new settlements of enterprises, services and companies whose activities are closely connected with the HT.

6. PEOPLE

Investigators operating at the Human Technopole will be:

- *fulltime* HT staff members, such as:
  - tenured scientists;
  - tenure track scientists;
  - researchers;
  - technologists;
  - postdocs;
- *a few* scientists with Joint Chair Positions between HT and Universities;
- *a few* scientists with Joint Positions between HT and other research institutions.

Associate members of HT would be

- university faculty/researchers (associate members);
- researchers from other research entities (associate members);
- researchers from industry;
- external collaborators;
- PhD students.

At steady state (i.e. after more than 7 years), the Human Technopole should involve approximately 1500 people (1000 staff + 500 PhD students). The staff distribution should follow the ratio shown in Table 1.

	P.I. (Directors and Tenure Track scientists)	**Researchers Technologists**	**Postdocs**	**PhD Students**	**Technicians**	**Administrative Staff**	**Support Staff** (Patents, Projects Office, Tech Transfer, ICT, etc.)	**Total**
Number	100	150	400	500	100	150	100	1500
Percentage	6.7%	10%	26.7%	33.3%	6.7%	10%	6.6%	100%

Table 1: Forecast distribution of HT Human Resources at steady state (> 7 years).

- The PI scientists with independent budgets and full scientific autonomy who coordinate Centers, Facilities, and/or Laboratories must be **full-time HT scientists**,. These scientists must be hired by international recruitment procedures. The positions can be Tenure Track or Tenured.
- Researchers and technologists are full time staff members acting as group leaders, staff scientists as well as facility scientists. They are hired by international selection, with time-limited contracts (typically 5 years);
- Post docs are recruited by standard international calls;
- PhD students in various disciplines will be enrolled through framework agreements with the Universities;
- Support staff will include experts in intellectual property and patents, technology transfer, technical office, scientific evaluation, research and project management, dissemination, outreach, and communication;
- Support, administrative, and technical staff will be recruited by open calls.

Salaries must follow European standards and will include a variable portion that depends on measurable results for **PIs, researchers/technologists, and managers** (i.e. MBO-Management by Objectives, evaluated by a Scientific Committee in the case of PIs). Details about the international recruiting system are given in Appendix 2.

Table 2 displays the planned growth of the Technopole's scientific and technical headcounts in the first 7 years, i.e. when steady state should be reached. These values are preliminary estimations and may vary during the recruitment process. However, they provide a reasonable indication of Technopole staff composition during the rump up phase. Note that Table 2 does not include: the associated researchers from the partner Universities and research institutions operating in the Joint Laboratories and the Seed Project Initiative.

Hiring Pgrm				**Total**				
Totale	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Total
Coordinators	7	-	-	-	-	-	-	7
Tenured / Tenure Track	1	27	29	13	11	9	2	92
Researchers	2	22	32	24	11	11	1	103
Technologists	-	9	10	1	1	1	-	22
PostDocs	8	70	175	61	64	47	10	435
Technicians	3	30	33	16	15	15	4	116
Subtotal - 1	21	158	279	115	102	83	17	775
PhD Students	-	75	139	100	63	48	14	439
Subtotal - 2	21	233	418	215	165	131	31	1.214
Admin & Support	19	48	52	50	28	27	26	250
Total HT								**1.464**

Table 2: Human Technopole's staff headcount in the first 7 years (from Rump up phase to Steady State). The values do not include outreach and University staff operating at the Joint Laboratories).

Table 2 assumes that each PI will form his/her own team within 24 months. The teams are standardized between 6 and 12 people (including PhD students, postdocs, technicians, and staff researchers) depending on the PI's seniority. According to this simplified model, the headcount should approach steady state after approximately 7 years, approaching **1500 headcounts** with the distribution shown in Fig.6.

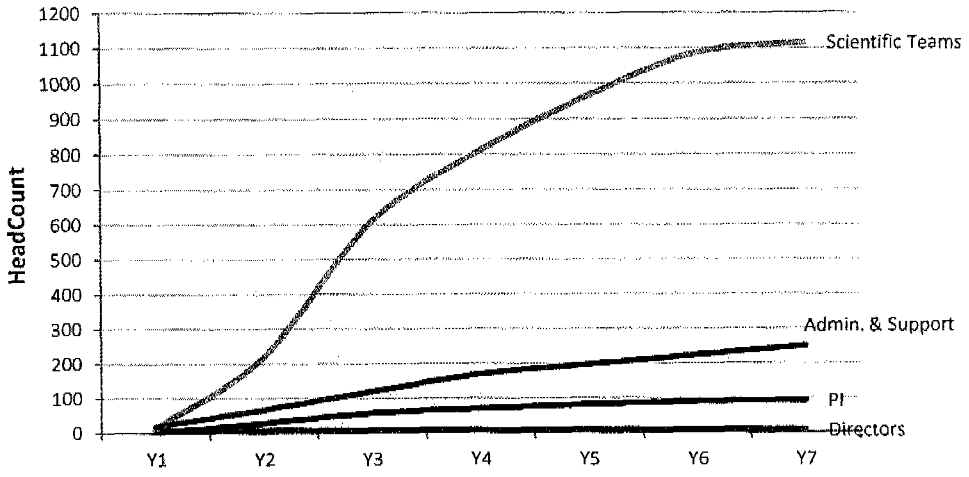
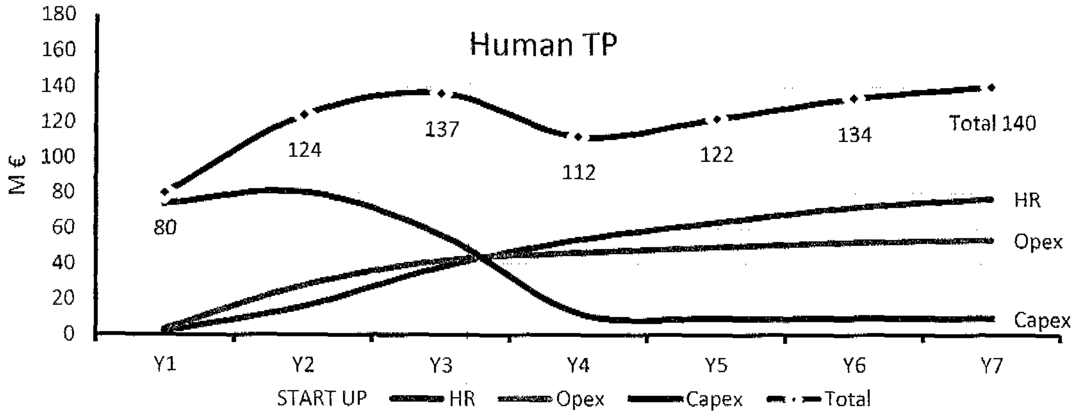


Fig.6 Foreseen hiring rate of the Human Technopole approaching steady state.

7. EXPECTED FINANCIAL NEEDS

It must be pointed out that a dedicated long-term funding law is required to ensure the HT's long-term sustainability. Based on the above forecast, Figure 7a displays the entire Human Technopole's financial needs in the first 7 years. Refurbishing and adapting the buildings are the most important items for the first two years. Staff and running cost expenditure are expected to grow continuously following international recruitment, overtaking capital expenditure in year 3. Approaching steady state, after year 7, the total financial need (including the Seed program) reachs about €140 million (44% of which dedicated to research running costs – CAPEX and OPEX- and about 56% to people). Figure 7b displays the headcount growth versus the financial need in the first 7 years. Steady state is approached after this period with a total budget of approximately €140 million/year and headcounts of 1464 units. At regime one can assume that the **average full cost per headcount stabilizes in the range of €95000 per year.**

a)



b)

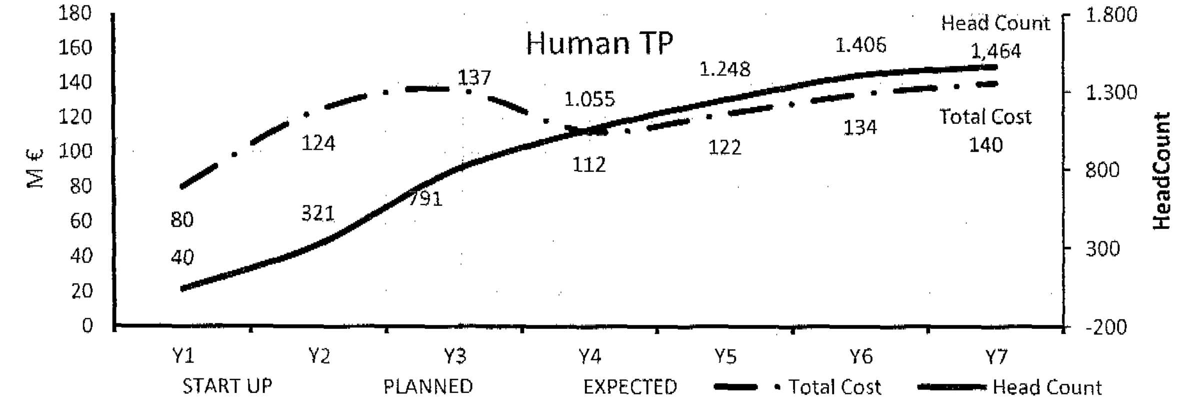


Figure 7 Forecast of the HT costs (M€) and headcount growth during the first 7 years; a) green, blue, and red lines indicate personnel, operational expenditures (OPEX), and capital investment (CAPEX, laboratory equipment, adaption/refurbishment of buildings) respectively. The black line displays the total; b) the black and red lines indicate the total cost and the headcounts, respectively.

Fund raising will be a primary target of HT. At steady state we envisage that HT will be able to raise up to 40% of its full cost. Competitive fund raising channels include Horizon 2020 calls, ERC and other individual awards, industrial grants and sponsored research agreements, charities. We also expect patent licensing and IP related revenues (e.g. royalties) to contribute to the global fund raising of HT. In addition to fund raising, indirect impact on GNP and social cost should be considered, such as start up creation, creation of new jobs, optimization of the public health system, etc.. The impact of the HT program on the health system, food and nutrition quality, predictive models for social needs and decision in the long term is indeed expected to be remarkable. Considering that in Italy the social cost of Cancer and Neurodegenerative diseases amounts to about 2% of the GNP (> 30 Billion Euro) per year, even a partial success of the precision medicine strategy would by far compensate the investment of the first 10 years of HT. A more detailed analysis of such forecast is provided in Appendix 3.

PART 2 TECHNICAL ANNEX

1. INTRODUCTION

Part 2 describes the Human Technopole's scientific case, the Research Lines to be developed by each Center, and the collaborative network inside HT and with external entities (the so-called "Interactome").

As discussed in the Executive Summary, the Human Technopole's key action in the start-up phase will be to recruit around 40 PIs (including the Directors of the seven Centers) via international calls together with around 60 additional PIs in the following years. The description of the scientific programs (Research Lines and Activities) is thus kept relatively general, defining the Human Technopole's global masterplan while still granting full autonomy to the new PIs to shape their own programs and teams within its framework.

Structure and joint collaborations: Each Center will comprise a headquarter at the Human Technopole, where Joint Laboratories with the Universities will be established, and several Outstations outside the Technopole (at major Research Institutions and Research Hospitals). The Joint Laboratories and Outstations displayed in Table 1 will should be established **during the start-up phase** via specific interinstitutional agreements (see also Fig. 2 of Part 1). Other Outstations and Joint Laboratories may be created during a more advanced phase.

	OGC	*NGC*	*AFNGC*	*DSC*	*CLSC*	*CADS*	*CSMD*
IIT		▨	▨		▨		▨
UniMi	▨	▨	▨				
Bicocca		▨			▨		▨
PoliMi						▨	▨
Humanitas	▨	▨					
IEO	▨						
INT	▨						
INGM	▨						
OSR	▨	▨					
Besta		▨					
M. Negri		▨					
FEM			▨			▨	
ISI				▨			
CREA			▨				
PTP			▨				
Cineca	▨	▨			▨	▨	

Table 1: The HT collaborations network during the start-up phase.

The OGC (C1) will comprise a central Research Facility at the EXPO site, incorporating several Joint Laboratories and Outstations. Collaborative agreements will involve the Universita' di Milano (*Clinical Genomics, Therapeutics and Clinical Trials*), Humanitas (*Immunogenomics)*, IEO (*Radiogenomics and Epigenomics*), INGM (*Functional Immunomics*), Ospedale San Raffaele (*Clinical Genomics*), and the Politecnico di Milano (*Computational Genomics*).

The NGC (C2) will comprise a central Research Facility at the EXPO site with Joint Laboratories and Oustations. It will include IIT (*Genomics, Bioinformatics, NeuroDrugMap, RNA-based Therapy, Neurodelivery*), l'Universita' di Milano (*Mesoscale Lab*), l'Universita' Bicocca (*Biomarkers*), Humanitas

(*Cognition and Plasticity, Neuroinflammation*), Istituto Farmacologico M. Negri (*Mouse Clinic*), and Ospedale San Raffaele (*Human Technologies*). The Istituto Neurologico C. Besta (with an associated network of Clinical Neurosciences) will direct the *Clinical Units* and the construction of the *Brain Bank*.

The AFNGC (C3) will comprise a central Research Facility at the EXPO site with Joint Laboratories and Outstations. It will include the University of Milano (*Smart Crops and Smart Food*), FEM (*Genomics and Post-Genomics of Fruit Crops*), CREA (*Seed Genomics, Precision Agriculture),* IIT *(Precision Agriculture, and Innovative Packaging*), and PTP (*Bioinformatics and Crop Biotechnology*).

The DSC (C4) will be headquartered at the Expo site, with an Outstation at the ISI Foundation. It will cooperate with the Università Statale di Milano, Università Bicocca, and Politecnico di Milano.

The CLSC (C5) will comprise a central lab at the EXPO site, with collaborative programs with IIT (*RNA Modeling, Genome Bioinformatics, Structural Genomics, Computational Drug Discovery*), Università Bicocca (*Computational Nanomedicine and System Biology*), and Politecnico di Milano (*Construction of the Storage/HPC Facility*). The partnership with CINECA will be crucial for the F3 Facility's data storage and HPC infrastructures.

The CADS (C6) will be jointly run by Politecnico di Milano in collaboration with FEM. It will rely on the shared Facility for Data Storage and High-Perfomance Computing (with CINECA).

The CSMD (C7) will be sinergystic with the existing IIT center for Nanoscience and Technology located at Politecnico di Milano with programs and joint laboratories with Universita' di Milano (*Human Sensing*) and Università Bicocca (*Nanotechnology for Food and Human Health, Valorization of Natural Polymers, Food and Agricultural Residues*).

Table 2 outlines the Center-Center interactions in the first years (Interactome). The green and yellow cells indicate experimental activities and computational activities, respectively. Collaborations and interactions between Centers will be supported by periodic internal calls for cross-disciplinary programs to be developed jointly by teams from different Centers. These will undergo international peer evaluation.

	OGC	NGC	AFNGC	DSC	CLSC	CADS	CSMD	F1	F2	F3
OGC	■	■	■	▨	▨	▨	■	■	■	▨
NGC	■	■	■	▨	▨	▨	■	■	■	▨
AFNGC	■	■	■			▨	■	■		▨
DSC	▨	▨		■	▨	▨		▨	▨	▨
CLSC	▨	▨		▨	■			▨		▨
CADS	▨	▨	▨	▨		■				▨
CSMD	■	■	■				■			
F1	■	■	■	▨	▨			■		
F2	■	■		▨					■	
F3	▨	▨	▨	▨	▨	▨				■

Table 2: "Interactome" of the Human Technopole. Acronyms stand for: OGC: Onco Genomics Center; NGC: Neuro Genomics Center; AFNGC: Agri Food and Nutrition Genomics Center; DSC: Data Science Center; CLSC: Computational Life Sciences Center; CADS: Center for Analysis, Decisions, and Society; CSMD: Center for Smart Materials and Devices.

Finally, periodic calls for ideas (Seed Projects) could be launched at the national level to support research programs and new ideas relevant to the Human Technopole's vision. In this way, the network of Institutions and teams contributing to the Technopole's activity will be expanded to include the rest of the country.

In what follows we provide the scientific description of Centers and Facilities.

C1: ONCO GENOMICS CENTER (OGC)

Vision: Genomics is at the forefront of innovation in biomedical sciences, transforming medical practice. Genomics is expected to critically contribute to the implementation of Precision (or Personalized) Medicine. This ongoing revolution in medicine (and life sciences in general) is likely to have an unprecedented impact on health and industrial development. This impact is already evident in oncology, where numerous gene-targeted therapies have been incorporated into routine clinical practice. Oncology is thus the initial choice for testing the near-term effects of Precision Medicine's implementation.

The application of genomics in oncology has revealed the challenges posed by Precision Medicine. These include the low number of tumor types for which there are targeted treatments with proven efficacy, the low number of eligible patients actually accessing the targeted treatments available, the difficulty of transferring to patients the continuous advances of genomics research (and thus extending the benefits of Precision Medicine), and the high costs of treatments and development pipelines. Precision Medicine is still seen as an extremely expensive specialized discipline that creates increasing costs for national healthcare systems, threatening their economic sustainability and their potential to improve public health. Thus, although genomics is expected to dramatically change public health management in the near future, it has yet to realize its radical potential. Finally, developments in medical genomics continue to reveal various ethical, legal, and social issues. Examples of these issues include the rights of access to and protection of personal genomic information, the need to provide patients and healthcare professionals with appropriate interpretative tools, and the difficulty in ensuring that the benefits and costs of genomics are fairly distributed across society.

Clearly, these issues must be addressed in order for Precision Medicine to realize its full potential in improving health and socioeconomic value. This will require a concerted mobilization of knowledge, expertise, and financial resources that are beyond the capabilities of any single research center. However, within the framework of dedicated centers focusing on highly specific projects, it is also crucial to address medical genomics' multiple components and implications in an integrated way. While individual projects may vary widely in their design and scope, their outcomes will form a broad foundation for applying genomic medicine and for understanding its multiple effects. These outcomes will also serve as blueprints for large-scale dissemination nationwide. Thus, a dedicated center is needed to take a multidisciplinary and highly focused approach to the delivery and implications of medical genomics and Personalized Medicine. This center should be rooted in the wider scientific community and in the various elements of society, including industry, the health system, and individual citizens.

The OGC's *mission* is to create a National Reference Center in Onco Genomics that will lay the foundations to extend medical genomics' current benefits to all Italian cancer patients. It will seek to continuously expand the oncology applications of medical genomics by rapidly transferring the knowledge arising from fundamental research. By creating a knowledge-based environment, the OGC will develop genomics sciences and technologies and rapidly translate scientific discoveries into *clinical* and *industrial applications*. This will be coupled to a new organizational model designed to drive scientific discoveries towards clinical and industrial endpoints. Today's low-efficiency linear scheme (basic -> translational -> clinical research) will be replaced by the *integration and mutual dependency of all knowledge-generation levels*. Preclinical and early clinical development will become a fast-track endeavor that is central to the OGC's mission. This will allow real-time transfer of new knowledge to clinical and industrial pipelines, and guarantee patients early access to innovative diagnostics and treatments. This process will be backed by world-class *fundamental research* to reveal the molecular components, mechanisms, and processes that underlie normal physiology and diseases, and that could represent future therapeutic targets.

In addition, the OGC will collaborate with other Italian academic centers and pharma/biotech companies: i) to disseminate genomics technologies; ii) to promote educational programs, in collaboration with Italian Universities, for emerging genomics-based skills (for clinical scientists, clinical bioinformaticians, quantitative biologists etc.); iii) to promote genomics culture and awareness via effective dissemination programs involving a broad community of stakeholders (population, patients, patient associations, health professionals); and iv) to promote new models for the economic sustainability of Precision Medicine by creating healthcare databases and launching pilot studies on novel clinical protocols and cost-saving procedures.

The OGC will comprise a central research facility at the EXPO site, incorporating several Joint Laboratories and Outstations. Collaborative agreements will involve the Universita' di Milano (*Clinical Onco Genomics, Therapeutics and Clinical Trials Office*), Humanitas and INGM (*Immunogenomics),* IEO (*Radiogenomics and Epigenomics*), Ospedale San Raffaele (Advanced *Clinical Onco Genomics*),IIT Center for Genomic Sciences – Milano (Cancer genomics) and the Politecnico di Milano (*Computational Genomics*). The *Clinical Onco Genomics* program entails the creation of a multi-institutional network involving several Italian Cancer Research Hospitals, including since the beginning the IRCCs in the Lombardy region (IEO, INT, Ospedale San Raffaele, Humanitas).

Scientific Structure: There will be three main areas of development:

RL1 Clinical Onco Genomics: The RL1 activities will seek to improve disease prevention and treatment strategies by promoting nationwide genomics-based screening programs, innovative clinical trials, and the generation of prescription and analytical computational tools. A common goal will be to translate the available basic knowledge into patient benefits and industrial applications in a rapid and cost-effective manner. This will require a series of tailored "top-down" thematic projects, such as large-scale genomic screenings. Because of their size, ambition, and costs, these projects will need dedicated institutional resources and strong cooperation between the various institutional components. Different Principal Investigators (PIs; from HT and other centers in Italy) will be involved in designing and executing these programs. The programs will be conducted in the OGC by dedicated teams (mainly technologists).

RL2 Fundamental Research in Onco Genomics: The RL2 activities will support continuous knowledge generation and innovation. They will be based on cutting-edge investigator-driven fundamental research, also known as "bottom-up" research. This will be developed in research groups led by fully independent Principal Investigators (PIs) recruited through international calls and with a proven track record in securing competitive external funds. RL2 research will seek to elucidate fundamental disease mechanisms and identify mechanism-based targets to drive the discovery of new drugs. RL2 research will also seek to reveal genomic markers for disease prevention or patient stratification, continuously replenishing the top-down pipeline of the clinical genomics projects.

RL3 Technological Platforms: The RL3 activities will contribute to the development and nationwide dissemination of new genomics technologies and computational workflows. RL3 will ensure the ongoing acquisition of the latest state-of-the-art genomics technologies.

Overall, the OGC will be structured as a *compact and fully integrated research operation* with all dissemination pipeline components (fundamental research, technology development, preclinical and clinical development) fully embedded into the Human Technopole's high-content scientific and technological environment (computer sciences & engineering, data analysis, mathematics, statistics, biology, genetics, health sciences).

As discussed below, OGC scientists will collaborate with other HT Centers on specific research areas (e.g. big data analyses, advanced bioinformatics, microbiomics and nutrigenomics, social and economic impact of genomics). They will also collaborate closely with other national and international research laboratories, offering access to advanced genomics technologies, dissemination of know-how, and integration into top-down screening and clinical development programs. Finally, the OGC will interact closely with biotech and pharma companies. By investing in highly innovative, high-risk translational projects (see below) to bring potential molecular targets to the *proof-of-concept* level, the OGC will create a continuous pipeline of projects of prospective interest for industrial applications.

The below section describes the Clinical Onco Genomics and Fundamental Research in Onco Genomics activities. The activities relating to the " *Technological Platforms* " will be described within the Central Genomics Facility F1.

Research Lines and Activities:

RL1 Clinical Onco Genomics.

Vision. One of Precision Medicine's central goals is to identify biomarkers that can be used for disease risk prediction, early detection, treatment selection, and treatment outcome monitoring. Indeed, the most effective targeted drugs are linked to biomarkers that predict treatment responses. The number of drugs with associated stratification biomarkers is rapidly growing. Within the large-scale omics technologies, genomic markers currently have the greatest impact. This is because next-generation sequencing (NGS) technologies are available at relatively low cost. These allow genome-wide analyses of DNA, RNA, and chromatin. To date, patients have been stratified using single genetic markers. However, Precision Medicine is now leading the transition to omics-scale diagnostics, including transcriptomics, epigenomics, proteomics, metabolomics, and lipidomics approaches.

However, several challenges considerably hinder the translation of Precision Medicine's rapidly expanding knowledge and tools into clinical practice. First, there are no standardized omics approaches for clinical use. Moreover, in terms of money, time, and human effort, the resources required are currently unsustainable in a routine clinical setting. These limitations significantly restrict patients' access to and benefits from personalized approaches. A scaled-down approach would improve accessibility enormously, but intensive efforts are required to optimize and standardize the trade-off between information loss and feasibility. Additionally, limitations in screening capabilities, drug availability, and practitioner training all restrict the chances of correctly assigning the best treatment given the underlying molecular defects. Finally, analytical approaches are heterogeneous and always changing. New strategies may offer improved results, but these must be correctly implemented to achieve measurable benefits. A useful example here is the analysis of the biological and genomic heterogeneity of normal and tumor tissues.

The *mission* of the OGC's Clinical Genomics Line is to implement genomic medicine in oncology in order to accelerate the transition to Precision Medicine. *Activities* will focus on the Actionable Genome, i.e. genes whose mutations in tumors predict specific drug treatments, individual allelic variants that predict cancer risk or drug toxicity, and molecules selectively present in cells of the tumor microenvironment that predict specific immunotherapies. The *General Goals* are: i) the *immediate clinical implementation* of the known Actionable Genome to improve cancer prevention and treatment for all Italian cancer patients; and ii) the *expansion of the Actionable Genome* by continuously incorporating newly discovered omics biomarkers into clinically relevant screening assays. The main *approach* will be to promote highly focused pilot research programs in genomic medicine nationwide. These may include broad genomic sequencing, integration of genomic data with clinical and personal data, and genomics-oriented clinical trials. These pilot programs will build capacity nationwide and demonstrate the effectiveness of these approaches before a full-scale implementation. In parallel, *another OGC goal* is to train a new generation of health professionals to be experts in genomics technologies (technologists) and/or in processing and using genomics and clinical data (i.e. clinical bioinformaticians and molecular pathologists with specific expertise in genomics). These activities will be strengthened by interactions with research lines at other HT centers (e.g. studies of population genomics within CLSC).

Clinical implementation of the known Actionable Genome.

Promotion of Nationwide Genomics Screening Programs for Patient Stratification.

This will be achieved in two ways. Firstly, the molecular profiling of cancer samples will seek to identify gene mutations that predict sensitivity to targeted drugs, including approved drugs and drugs undergoing clinical trials (Somatic Actionable Genome; currently a few hundred genes). Secondly, the molecular profiling of the DNA of individuals with suspected familial susceptibility to cancer will seek to identify germline variants that predict cancer susceptibility (Germline Actionable Genome; currently a few hundred genes). The medium-term vision is to offer molecular profiling to all Italian cancer patients and, if familial cancer is suspected, their relatives. The project will be implemented in stages, beginning with the most frequent types of cancer (breast/ovary, prostate, lung, colon) and involving selected Italian Research Hospitals (see below). Whole Genome and Whole Exome Sequencing will likely still be too expensive for a population-wide screen. An initial approach might therefore be to sequence selected genes (gene-panel NGS analyses). By designing specific gene panels and developing a robust sequencing technology, the aim is to identify all genetic variations (SNV, CNV, translocations), while continuously upgrading the design for an inexpensive implementation (a few hundred euros per sample).

Promotion of Genomics-based Clinical Trials.

The clinical value of the currently known Actionable Genome is very limited. After profiling of the publicly available cancer genomes (~15,000 samples), in silico prescription of new drugs predicts that less than ~10% patients can benefit from approved targeted therapies. If we consider the pipeline of new drugs in clinical trials, the repurposing of approved drugs, and drug combinations, the percentage of patients who could benefit from genomics-based treatments rises to ~70%. Thus, genomics-based clinical trials (especially those led by academia) can reveal how newly targeted drugs can be most effective. These clinical trials can also expand and standardize genomics medicine applications, guaranteeing patients early access to oncology innovations. Genomics screening programs and genomics-based clinical trials must thus be promoted in an interconnected manner to maximize their benefits to patients. However, genomics-based clinical trials have unique limiting features. These include the extreme fragmentation of the patient population and the availability of powerful biomarkers of response. Addressing these features requires multicenter academic clinical trials, the execution of complex genomic analyses, and the design of innovative clinical schemes. Within HT, a primary objective is thus to create a structure (the Clinical Trial Office) that provides cutting-edge expertise (clinical statisticians, medical oncologists, experts of regulatory issues, bioethicists) to Italian Research Hospitals involved in genomics-based clinical studies. This will promote a culture of innovative genomics-based clinical trials, facilitating their design and implementation. It is mandatory that the HT Clinical Trial Office collaborate with the Ministry of Health, Regional Governments, and AIFA (Agenzia Italiana del Farmaco, Italian Agency for Drugs) to discuss strategies for promoting, regulating, and funding genomics-based clinical trials. We hope this collaboration will result in novel standard operating procedures being developed for the design of genomics-based clinical trials.

Generation of Large-ccale Genomic and Clinical Data Resources (Prescription and Analytical Computational Tools).

The handling of patient data is a question of paramount importance. For each patient, a huge amount of "personalized data" must be collected and integrated (including omics, environmental, and lifestyle data, medical history). Furthermore, to be clinically relevant, each individual patient's genomic data must be linked in real time to the currently available scientific information (i.e. the information that defines the data's clinical actionability). The scale of the emerging data is thus gigantic and outpaces our human cognitive capacity. We will undoubtedly require extensive computational infrastructures and pipelines to organize these data (large-scale genomic and clinical data resources), to analyze new genotype/phenotype correlations (analytical tools), and, for each single patient, to facilitate the interpretation of their genomics data and the consequent medical decisions (decision-support tools). A number of databases must be built, including a Mutational Database and an Actionability Database. The Mutational Database will contain genomic data from patients and be built dynamically with different levels for extracting the genomic information. To increase the population sizes for exploratory analyses, the Mutational Database should be linked to publicly available genomic data. The Actionability Database will contain all the information on genomic actionability, starting with the available public resources. This database should be equipped with semantic structure extraction algorithms for machine-assisted literature curation. Both databases should be linked to hospital databases including the EMRs (Electronic Medical Records), to Regional Healthcare Databases (drug prescription, hospital discharge forms, outcome monitoring, etc.), and to Research Pipelines (advanced genomics, epigenomics, and proteomics technologies). However, these data are heterogeneous and stored in different formats. Therefore, adequate semantic mapping must be implemented to order information into a uniquely integrated data structure that can support efficient querying. Integrating all these databases will provide pipelines that can support clinical decisions for individual patients, allow cross-patient analyses, and couple genetic and clinical information to clinical trials. This crucial task is immensely complex and will require many additional complementary skills (complex database construction, network architecture, data mining, and interface design). In view of the high concentration of resources and know-how in data analysis and IT infrastructures at its disposal (particularly within F3, C4, and C5), the HT offers a unique opportunity to accomplish this goal.

Institutional Context.

Implementing national genomic screening programs for patient stratification, promoting genomics-based clinical trials, and generating prescription and analytical computational tools can only be achieved within a National Network, in close collaboration with selected Italian Research Hospitals and healthcare governance organizations. Similar initiatives have already begun in several Italian institutions but, being highly fragmented and lacking adequate funds, they risk the dispersion of resources and expertise. This project's success depends on establishing a multi-institutional network involving several Italian Research Hospitals. This network can thus achieve a critical mass of complementary expertise, adequate patient cohort sizes, and an efficient and cost-effective sharing of advanced technological resources. In addition, from a scientific point of view, multicenter and cooperative research projects are necessary to address the fragmentation of the cancer patient population as a consequence of genomic stratification.

Within this network, the OGC will act as a central facility of sequence capabilities and data/sample storage. It will also build methodologies and workflows with clinical scientists, transferring know-how, sequence capabilities, and computational tools to the network's Research Hospitals.

This will require: i) networking the preexisting DNA-sequencing facilities and genomic technology expertise; ii) setting up common experimental protocols and bioinformatics pipelines to clinically validate the available genomic markers; and iii) integrating the generated genomic data with clinical information in a shared database. Peripheral centers can then access this database, which will play a central role in their day-to-day clinical care. We envision a similar level of integration with the National Healthcare System in terms of access to Regional Healthcare Databases.

This complex coordinating taskforce could be strengthened by a formal joint venture with Alliance Against Cancer (ACC). AAC is an Italian network of 19 Research Hospitals (IRCCS) involved in cancer research and treatment. It is coordinated by the Ministry of Health. In particular, given their physical proximity to the Technopole, the ACC IRCCs in the Lombardy region will be involved from the very beginning.

The ACC network has the advantage of being distributed throughout the country. It is also big enough to support a national screening pilot project and pilot clinical trials involving cancer patients (~ 90,000 new cancer patients per year and 70,000 patients in clinical trials). Furthermore, the network is already working to set up a national genomic screening project for patient stratification. Added value will come from the involvement of the Ministry of Health and Regional Governments. Their involvement will extend the project's goals to include economic endpoints (financial sustainability) and test models of governance of innovation in oncology.

Ethical, Social, and Legal Aspects (in Collaboration with CADS).

On a strictly technical level, in order to conduct national screenings and promote genomics-based clinical trials, specific consent forms must be prepared to secure patients' informed approval. The set-up of a healthcare database must also comply with rules safeguarding patient privacy and patient rights. Unfortunately, the current rules do not cover many of the ethical and regulatory issues related to genomic medicine and big data in the field of healthcare. An impasse could thus arise from the apparent contradiction between the confidentiality of genomic information and the use of this data to improve public health. The following are examples of issues that will need to be addressed: patient privacy and confidentiality; data control by single individuals; data ownership and control over the transparency of ownership; the implications of data analytics for personal privacy; patients' capacity for interpreting and managing data; doctors' understanding of patients' behavioral responses to genomic medicine; and the effects of the ongoing developments on data control (crowdsourcing, participatory surveillance) and data acquisition/sharing (social media, GPS-enabled mobile apps, and tracking/wearable devices). This part of the project will be led by the HT Center for Analysis, Decisions, and Society (C6).

Economic Aspects.

Various data items are needed when designing studies to economically evaluate the various medical genomics activities. Establishing a Health Database is a critical opportunity to gather these items. This part of the project will be led by C6 (HT Center for Analysis, Decisions, and Society).

Expansion of the Actionable Genome.

Set-up of Large-scale Patient Genomics Screenings to Identify New Stratification Markers (in the Order of Thousands of Samples).

DNA sequencing (whole genome or whole exome) of a large number of cancer genomes has provided new treatment stratification markers and shed new light on tumorigenesis mechanisms. Today, around 15,000 cancer genomes are available on public databases, with many massive cancer-sequencing programs under way worldwide. It is likely that sequencing the entire spectrum of cancer-associated relevant mutations will be completed very quickly, even for less frequent tumor types. We intend that the OGC's high-throughput sequencing efforts will focus on subsequent applications, for which the available data are not adequate. One current critical issue is predicting a tumor's sensitivity to new therapeutic approaches (most notably immunotherapy) and, in general, its mechanism of resistance to new drugs. Screenings of this kind are feasible on a more limited number of patients (in the orders of a few thousand) and within collaborative studies, where cancers are sequenced before and after specific treatments. Such studies require a high degree of integration with fundamental genomic research activities. For example, when screening the immunogenic genome, cancer-genome analyses should be integrated with analyses of the predicted immunogenicity of cancer mutations, with functional genomic studies, and with detailed molecular descriptions of the cellular components of the tumor microenvironment.

Set-up of Large-scale Genomics Screens to Identify New Cancer-Predisposing Genes (in the Order of Thousands of Individuals).

The type of genetic screening used to date (for example, linkage or candidate-gene analyses, GWAS) has identified only a portion of the genetic risk factors (e.g. rare high-penetrance genes and common low-penetrance variants). This suggests that most of the genetic risk has yet to be discovered (e.g. if mediated by a large number of low-frequency moderate-penetrance genes). Sequencing the entire coding region in the human genome will likely help identify new genetic factors for cancer susceptibility. Such studies can exploit well-characterized cohorts from large epidemiologic studies. A useful example is the EPIC cohort study, which recruited ~ 500,000 participants in 10 Western European countries, including Italy, and followed them for almost 15 years for disease endpoints, vital status, and causes of death (~ 100,000 cases of cancer). It is envisioned that HT will collaborate with such groups to perform genome screenings of highly selected populations in order to identify novel genes and pathways involved in genetic susceptibility to cancer in Italy. Using these kinds of populations may also provide an opportunity to explore the relationships between genes, diet, nutritional status, lifestyle, and environmental factors.

New approaches to investigate the biological and genetic heterogeneity of tumors.

Recent discoveries have shown an unexpected genetic and biological heterogeneity of both normal and cancer tissues. In fact, both contain functionally distinct cell types and multiple genomes, some at very low frequency. Tissue heterogeneity has important clinical implications for the risk of cancer development (e.g. emergence of cellular clones within normal tissues) and the response or resistance to new drugs (e.g. selection of preexisting rare mutations in tumor tissues). Tissue heterogeneity must thus be thoroughly investigated. To enable these investigations, new technologies must be established and progressively incorporated into the current pipelines for screening cancer patients. These technologies could include: *a)* ultra-sensitive sequencing protocols (to identify low-frequency mutations in the bulk tissue, far below the current limits of ~10%); *b)* single-cell genomic analyses (to explore genetic variability directly at the single-cell level); *c)* single-cell phenotypic analyses (to identify and characterize rare cell populations); and *d)* single-molecule/long-read sequencing (to identify combinations of genetic/epigenetic variants and structural variants). These technologies will be established in collaboration with basic scientists, who will tackle the same issues with a mechanistic approach and using model systems (see *"Common Technology-Development Platforms"*). It is crucial, however, that these technologies are applied to and optimized for patient samples, including cancer tissues, circulating cancer cells, and the different body fluids during their development.

Beyond Genomics.

The aforementioned Actionable Genome is based on the genetic alterations found in cancer cells, as determined by genome sequencing. The Actionable Genome is one route to intervention. Another route is provided by mechanistic alterations that are not caused by direct mutations, but rather by regulatory changes, such as epigenetic events (e.g. DNA methylation, histone modifications) or imbalances in transcription factor networks, which can cause overexpression or silencing of important genes (e.g. oncogenes or tumor suppressors). We can refer to these events collectively as the "Actionable Epigenome." The Actionable Epigenome is another fundamental but highly complex source for the development of anti-cancer therapies. These events cannot be detected by direct sequencing; however, understanding and exploiting their complexity involves basic mechanism-oriented research efforts, which rely in part on the same high-throughput sequencing technologies, in particular for determining gene expression (e.g. RNA-seq) and epigenomic profiles (e.g. ChIP-seq). A similar systematic effort could also be extended to include the proteome (through analysis of protein levels and protein post-translational modifications) as well as changes in levels of metabolites and lipids (metabolomics and lipidomics).

The OGC's activities in this sector will follow two strategies. First, the Clinical Genomics program will undertake projects to identify mechanism-based targets, expanding the clinical development of the Actionable Epigenome, just as it will do for the Actionable Genome. Second, basic research will aim for a deeper mechanistic understanding, which is essential to drive clinical development. This is the essence of the OGC's Fundamental Research in Onco Genomics, described in the next section.

RL2. Fundamental Genomics.

Vision. This research line's main goals are: i) to perform innovative genomics research, focusing on cancer-associated disease mechanisms, in order to identify molecular targets and markers for prevention, early detection, and personalized treatments; ii) to disseminate theoretical and technical expertise in genomics to Italian Research Institutions.

There are five leading priorities: *Cancer Genomics, Epigenetics and Transcription, ImmunoGenomics, RadioGenomics*, and *Therapeutics*. The underlying vision is briefly described below. These areas are complementary and largely overlap in terms of objectives and technological/experimental approaches.

Cancer Genomics.

The advent of NGS has brought unprecedented insight into the transcriptional, epigenetic, and mutational landscapes of cancer genomes. Previously, the focus was on the identification of mutations in single cancer genes and on the molecular dissection of the underlying signaling pathways. Now, NGS-based technologies provide a full-scale view of gene mutation combinations and their heterogeneity in cancer genomes. These studies have also highlighted new layers of complexity. These include: mutations/alterations of regulatory regions and in repetitive DNA; local/global epigenetic alterations and in the nuclear topology of cancer (epi)-genomes; and alterations in RNA/DNA editing of cancer genomes/transcriptomes. However, it is still unclear precisely how most cancer-associated genetic and epigenetic alterations contribute to the disease. This has prompted calls for a fundamental research effort focused on *disease mechanisms*. The resulting gain in knowledge will, in turn, accelerate the move to fully realizing the Precision Medicine concept, optimizing the risk assessment, diagnosis, and treatment of cancer patients.

The systematic sequencing of individual genomes and epigenomes must be complemented by advanced *preclinical models* that provide accurate disease replicas for research and therapeutic development. Currently, patient-derived xenotransplants (PDX) are limited by, for example, the lack of a humanized and fully competent immune system. These limitations must be overcome by more advanced models (PDX in immunodeficient mice reconstituted with human autologous or heterologous hematopoiesis). In parallel, ex vivo systems (tumor organoids) must be set up to reconstitute the 3D structure and microenvironment of tumor tissues, and to perform genetic and pharmacological screenings in a time-efficient and cost-efficient manner.

These models will allow *functional genomics* approaches and facilitate detailed mechanistic studies of the molecular and cellular basis of this disease. The models will also allow sophisticated assays suitable for testing in collaboration with other HT centers (e.g. novel nanomaterials in collaboration with CSMD, other activities described below). Taken together, the above research activities will elucidate how specific alterations – from among all those occurring in cancer cells - influence the onset, progression, and maintenance of cancer. This should help identify novel molecular markers (or *biomarkers*) and *targets* with important diagnostic, prognostic/predictive, and therapeutic implications. Dedicated activities will be required to translate selected biomarkers and targets into clinical practice (see *Therapeutics* section below). The two research areas with the greatest potential to positively impact cancer patients are: i) innate and adaptive immune responses to tumors and ii) epigenetic and transcriptional mechanisms of tumor development. Microbes, food, and their impact on genetic regulation will also be systematically investigated to reveal their impact on cancer and related therapies (in collaboration with NGC and AFNGC, see below for a description).

Epigenetics and Transcription.

Transcriptional control and chromatin-mediated regulation of the transcriptional machinery's access to the underlying genome are at the basis of all normal and pathological cellular properties. These processes are frequently altered in cancer and are thus potential targets for highly innovative drugs.

In normal developmental processes, sequence-specific transcription factors (TFs) and transcriptional co-regulators enforce the creation of distinct cell types using the same genome. This diversification is regulated by distinct signaling pathways. It involves the selective activation of distinct fractions of the genomic *cis*-regulatory information, leading to the establishment of transcriptional outputs that are unique to a given cell type and functional status. Chromatin modifications associated with differentially active genic or cis-regulatory regions are like footprints left by the activity of TFs and the macromolecular machineries. As such, these modifications bear information that can be decoded to interpret the cis-regulatory genome and to map signals upstream.

In *cancer*, the same machineries are altered by mutations of their components and by aberrant signals, which can originate from genetic alterations in signaling pathways, or from the tumor microenvironment. Consequently, cancer cells activate transcriptional networks and cellular programs that differ to varying extents from those of normal cells, giving rise to distinct tumor types. Within this conceptual framework, the new properties acquired by cancer cells reflect reprogramming events. Tumor cells should thus be considered as novel and distinct entities, with stable and unique gene regulatory networks and chromatin patterns. Deconvolution of the regulatory alterations in cancer genomes may thus help identify cancer-relevant signaling pathways, which may themselves be novel targets for therapeutic interventions.

Research in this area will include basic and preclinical studies to *elucidate normal and pathological gene regulatory networks* whose components could be *targeted for therapeutic purposes.* Molecules involved in transcriptional control have traditionally been seen as undruggable. However, recent advances in pharmacological targeting of protein-protein interaction surfaces have paved the way for next-generation novel drugs that can modulate transcription. The program will rely heavily on implementing *cutting edge omics as well as genome and epigenome editing technologies.* Its goal will be to obtain a broad unbiased description and extensive functional elucidation of the epigenomic, transcriptomic, and proteomic changes that occur in physiological and pathological transitions. In this research area, we will also study theoretical 3D models of genome organization developed by other HT centers (C5) and approachable experimentally. *Technological development* will also focus on *understanding cellular heterogeneity in normal and diseased tissues.* It will include the analysis of *single cells,* both isolated *ex vivo* and *within intact tissues,* by combining advanced genomics, proteomics, and imaging approaches (see "*Common Technology-Development Platforms*" within Facility F1). Finally, datasets generated within this research area and associated with patient-derived information will *feed large-scale data analysis efforts within the Human Technopole* (particularly within DSC and CLSC).

Immunogenomics.

Most pathophysiological processes, ranging from aging to metabolism, involve orchestration of the innate and adaptive immune systems. Inflammation controlled by innate immunity mechanisms is a key component of the tumor microenvironment, and taming of adaptive immune responses by cancer cells plays a key role in tumor progression. Harnessing the immune system against cancer and blunting tumor-promoting inflammation are strategies at the forefront of cancer treatment and have already produced great breakthroughs in recent years. In addition, immunity has emerged as a trans-disease mechanism, broadly involved in the pathogenesis of neurodegenerative, cardiovascular, and metabolic diseases. Within this diverse set of disorders, Precision Medicine calls for an integrated *omics*-based approach to deipher the full complexity, diversity, and plasticity of the immune system (immunogenomics) and its variation in human tumors. Through different "omics" (i.e. genomics, transcriptomics, epigenomics, metabolomics, and microbiomics), we expect immunogenomics to provide new insights into cancer and degenerative disorders, new tools for Precision Medicine, and a unique approach at the intersection between food, microbes, and health.

To achieve these goals, coordinated genomic, transcriptomic, and epigenomic approaches must be combined with analysis of immune system function and the clinical outcome of patient cohorts. Information and analysis must move between laboratory bench and hospital bed in both directions. Innovative preclinical models must also be developed. In particular, we plan to: *i)* investigate the commitment, plasticity, and regulation of immune cells that interact with tumor environment (i.e. tumor-infiltrating lymphocytes, macrophages, and other myeloid cells); *ii)* characterize at the single cell level the molecular mechanisms and networks underlying the innate and adaptive immune responses in cancer; *iii)*define from a molecular point of view "cancer immunoediting", i.e. the mutual influences exchanged between cancer cells and immune cells; and *iv)* produce novel models (*ex vivo/in vivo*) of immune system/tumor cell interactions and cross-talk (such as immune competent patient–derived xenograft - PDX models, or tumor organoids reconstituted with patient-derived immune cells).

Immunotherapy ranges from checkpoint inhibitors to cell therapy, including CAR and the blocking of tumor-promoting inflammation. It is emerging as a game-changer for anti-cancer therapy. However, to fully exploit its potential, cancer genomics and immunogenomics must be integrated to define, at the molecular level, subgroups of patients with specific therapeutic indications. The expected benefits include personalized precision-based treatments, the identification of new biomarkers, and the development of novel candidate-targeting strategies. Critical objectives include: *i*) identifying cancer patients responsive to immunotherapy by analyzing immune cells and cancer genomes in patients who are either responsive or unresponsive to immune-based therapies; *ii)* developing and/or characterizing novel markers and antibodies for therapeutic intervention; and *iii)* identifying new molecular targets through functional screens in new model systems (humanized models or organoids); iv) ideation, planning and execution of new pilot phase I clinical trials in which the best immunotherapies will be combined with other anticancer therapeutics approaches

RadioGenomics.

Radiotherapy is the oncology treatment with the single strongest impact on patient survival. It is thus critical to understand the genetic bases for differences in sensitivity to radiation therapy. We plan to increase understanding of individual tumor sensitivity to radiotherapy by using the currently available omics tools to explore the genetic make-up, cell heterogeneity and immune components of tumors. Studying the correlations between genomics and radiosensitivity will inform the set-up of more efficient radiotherapeutic strategies (increased tumor control and reduced toxicities) and healthcare plans (reducing the financial burden of treatment failures and/or toxicities). Critical activities in this area include *i*) characterizing the molecular mechanisms of radio-resistance in distinct tumor cells (e.g. cancer stem cells) in order to identify sensitizing targeting agents or particles; *ii)* defining the relationship between the cancer genome and the effects of new radiation modalities, such as high LET particles (helium, carbon, and oxygen ions) and low-dose exposure; and *iii)* developing risk models, including genetic factors and clinical covariants, to predict individual radiosensitivity and side effects after external radiotherapy. Molecular data will be integrated with tumor-imaging data (NMR, PET, and enhanced PET). The entire collection of datasets *will then feed large-scale data analysis efforts within the Human Technopole.*

Therapeutics.

The integration of genomics, proteomics, and functional genomics data will generate knowledge that must be translated into the better use of existing drugs, the identification of novel drug targets, and the generation of novel drugs and innovative treatments. Critical research activities will include: *i) Identification and validation of novel targets.* Novel targets must be identified systematically based on their functional relevance for specific cancer patients (as defined by their (epi)-mutational landscape). We will therefore perform genetic/pharmacological in vivo/ex vivo screenings in second-generation PDX models/tumor-organoids from thoroughly characterized patient-derived samples. The identified gene targets will be then mapped into gene networks to define targeting strategies that are "pathway-based" rather than "gene-based". Of the candidate targets, epigenetic and metabolic targets should be prioritized in terms of how they integrate with other activities/expertise within OCG or within the wider scientific community in Italy. *ii) Identification of novel chemical probes/drug prototypes against newly identified target genes/pathways.* High-throughput functional screens are expected to identify novel gene targets, for which there is no proof of principle of druggability (to date, this is mainly limited to enzymes). Small molecules that may function as chemical probes will be isolated using modern large-scale chemical-genomics studies and then used to: a) demonstrate the druggability of the identified gene targets; b) provide chemical tools for further validation in biological assays of the candidate target; and c) provide initial chemical scaffolds/hints for a more direct drug discovery approach. At this stage, we will also explore multivalent small molecules derived from rational design approaches in order to target cooperating genes within the same pathway. *iii) Identification of novel strategies for drug treatment, based on quantitative approaches.* Drug treatments must be optimized through the development of population dynamics quantitative models, keeping in mind the functional heterogeneity of tumor cells, and their continuous genetic and epigenetic evolution into different cell subpopulationsAt present, hybrid probabilistic/quantitative models appear the most appropriate to tackle simultaneously these aspects. Moreover, these models allow to highlight the potential emergence of drug resistant phenotypes and toxicity that may be efficiently minimized through the identification of the most effective treatment schedule. Treatment modality/schedules of drug combinations will be then validated in preclinical models, as the basis for final validation in clinical studies.

These activities will be fully integrated with the Human Technopole's CLSC, which will provide critical expertise and tools for the systematic deconvolution of genetic/drug screens and computational screenings. These will be used to identify chemical hits/probes for chemical genomics and initial drug discovery activities, as part of the CLSC research activities in structural genomics, computational drug discovery, and polypharmacology (see description of CLSC). To generate innovative drug discovery tools, the activities will also be fully integrated with a national and international network of academic collaborators in medicinal and computational chemistry. Particular emphasis should be given to collaborations with the Italian medicinal chemistry community, which is top-ranked internationally but not yet fully exploited. Finally, close interaction with the HT technology transfer staff will be integral to this research line. We will identify industrial and venture capital partners who can immediately transition to identifying new drugs. Notably, this research line does not seek to generate candidate drugs or undertake conventional drug discovery. Rather, by focusing on high-risk high-innovation targets, this research area seeks to generate chemical tools/probes to provide initial proof-of-concepts and to justify further drug discovery efforts. In our vision, these efforts will be pursued by nonacademic entities (venture capital/industry) in collaborative partnerships with HT. As such, this research area will address a key domain that is rarely exploited by industry (biotech/pharma), seeding innovative projects to be brought to fruition by industry.

Data Analyses.

The OGC activities will generate immense datasets that will *feed large-scale data analysis efforts within the Technopole.* An integrated systems approach will be needed to extract the relevant information and provide a more comprehensive view of the biological context. This will be the challenge for bioinformaticians, biostatisticians, and computational biologists tasked with developing novel solutions to improve data use and scientific discovery. These tasks will be accomplished in close collaboration with the DSC, CLSC, and F3 of HT.

C2: NEURO GENOMICS CENTER (NGC)

Vision: We aim to apply Precision Medicine to neurodegenerative diseases. Alzheimer's disease (AD), Parkinson's disease (PD), and Amyotrophic Lateral Sclerosis (ALS) are all incurable and their molecular mechanisms remain unclear. The list of failed clinical trials is discouragingly long and the amount of money spent is so high that pharmaceutical companies are considering abandoning this field. We believe this state of the art is due to several largely unsolved issues in basic research departments, current traditional clinical settings, and drug development strategies. The major issues are as follows:

- Most understanding of the pathways involved in neurodegenerative diseases has been derived by identifying genes mutated in familial cases. Such cases represent, on average, 10% of patients. Cellular and genetic animal models have been produced, but drugs that were neuroprotective in these models failed when tested on patients.
- There is no molecular-based classification of patients suffering from sporadic diseases. We still do not know the molecular signatures that distinguish patients from healthy subjects. More importantly, these diseases are not homogeneous. The lack of a molecular-based stratification can have a very negative impact on the outcome of clinical trials, where drugs may fail due to a limited number of patients with varied genetic backgrounds.
- At the onset of symptoms, many neurons have already been lost. This indicates that neurodegenerative diseases have a long presymptomatic phase, lasting several years. At the point where drugs are tested, neuronal networks have already been irreversibly compromised.
- The integrated searchable records of clinical and genetic data together with data on lifestyles and nutrition habits are, for the most part, very poor. They involve a small number of patients and are scattered throughout different institutions.
- To date, all the genetic association studies have focused on analyzing genomic DNA from blood, which ignores the potential role of somatic mutations.
- Most attention has been given to the exomes of protein-coding genes. These approaches have not addressed the potential role of regulatory elements, long noncoding RNAs, and repetitive sequences.
- All genomic analysis has been performed on a large number of cells, ignoring cell-to-cell variability.
- Because drug development is based on single-gene-mutation models, combinatorial drug treatments have been ignored.
- Drug repositioning has been very limited because there is no connectivity map of FDA-approved drugs for the nervous system.

To overcome these issues, NGC will carry out a coordinated effort to establish Precision Medicine for neurodegenerative diseases. This will be accomplished by: i) classifying patients according to their genomic sequences associated with a group of symptoms; ii) identifying biomarkers to diagnose disease; iii) monitoring disease progression; iv) deepening understanding of disease mechanisms, and v) developing new treatments.

While the spectrum of neurodegenerative diseases is very large, we will focus on AD, PD, ALS, and related disorders. These have been chosen for their staggering social impact, the lack of existing therapies to block or slow down neurodegeneration, and, given the Human Technopole's location, the possibility of using the expertise of the neurological institutions in the Milan area. Importantly, since these diseases may manifest overlapping molecular and neuropathological signs, it is crucial to study them in an integrated fashion.

The NGC will comprise a central research facility at the EXPO site with Joint Laboratories and Oustations. It will include IIT (*Genomics, Bioinformatics, NeuroDrugMap, RNA-based Therapy, Neurodelivery*), l'Universita' di Milano (*Mesoscale Lab*), l'Universita' Bicocca (*Biomarkers*), Humanitas (*Cognition and Plasticity, Neuroinflammation*), Istituto Farmacologico M. Negri (*Mouse Clinic*), and Ospedale San Raffaele (*Human Technologies*). The Istituto Neurologico C. Besta (with an associated clinical neurosciences network) will direct the *Clinical Units* and the construction of the *Brain Bank*.

Scientific Structure: The Center will be organized along 3 Research Lines *RL1 Neuro Genomics, RL2 Models & Mechanisms*, and *RL3 Translational Neuro Genomics,* together with the "Common

Technology-Development Platforms" in Facility F1 and the iPSCs platform in Facility F4. A single *RL* will contain several Research Activities carried out by one or more independent PIs (tenure-track scientists), who will lead a group of 5-10 PhD students and postdocs. International calls will be launched promptly to recruit PIs and to activate the research activities. NGC will be closely connected to the other Human Technopole Centers, in particular with AFNGC, CLSC, DSC, and CNST. Below, we describe the activities for *Neuro Genomics, Models & Mechanisms*, and *Translational Neuro Genomics*. The activities related to the "Common Technology-Development Platforms" will be described within the Central Genomic Facility F1 section and the iPSCs platform section in Facility F4. NGC will make it a priority to develop close interactions with biotechnology and pharmaceutical companies.

Research Lines and Activities:

RL1. Neuro Genomics.

Vision. Here we seek to establish a genomics-based stratification of patients with neurodegenerative diseases. This will be accomplished by enrolling patients and healthy controls to create a tissue bank, performing genomics/transcriptomics analysis, and assembling a searchable database with electronic records of patients' medical history, lifestyles, and nutrition habits together with data from in-depth clinical assessments. To achieve this, we will bring together a network of Clinical Neuroscience Centers. For each participant in the study, we will:

- collect and analyze biological samples. For living individuals, we will harvest several peripheral tissues including blood (plasma, lymphocytes, exosomes), dermal fibroblasts, and cerebrospinal fluid. For metagenomics studies of the microbiota, we will harvest patients' stools.
- sequence the whole genome from blood as well as transcriptomes of lymphocytes, exosomes, and cerebrospinal fluids. We will also sequence the gut microbiota.
- prepare and store iPSCs from each individual.
- carry out a detailed clinical assessment according to the expertise of the neurologists at the Clinical Neuroscience Centers.
- enter all the experimental and clinical data into the Human Technopole database, including medical history, lifestyles, and nutrition habits.
- perform an in-depth histopathological analysis of post-mortem brain tissues. Whole genome sequences of brain areas and/or single neurons will be carried out and compared to sequence assemblies from peripheral tissues to identify somatic variants.

RL1 will be carried out in collaboration with CLSC (Population Genomics, Structural Genomics, and System Biology) and DSC. It will take advantage of F1 for sequencing and F4 for iPSC preparation and storage.

Clinical Units (Outstation: IIT@Besta and Hospital Network).

The Outstation comprises five Clinical Neuroscience Centers, which are internationally renowned for their clinical and preclinical research on neurodegenerative diseases (particularly AD, PD and ALS) and for the availability of relevant cohorts of patients and biological material. This network's Technopole-related activities will be coordinated by the IRCCS Foundation "Carlo Besta" Neurological Institute and will be carried out by i) Carlo Besta Neurological Institute, ii) Ospedale Maggiore Policlinico, University of Milan, iii) San Gerardo Milano-Bicocca University Hospital, iv) Ospedale San Raffaele, and v) Humanitas Research Hospital.

The primary objective is to build up large cohorts of extensively characterized individuals with AD, PD, ALS, and related disorders, and to collect biological samples for genomic analysis and molecular classification, nutrigenomics, identification of predictive, diagnostic and treatment-responsive biomarkers, and patient-derived cells for mechanistic studies. Since a critical aspect of AD, PD, and ALS is their phenotypic heterogeneity and the clinical overlap with other neurological disorders, the Outstation will use the existing advanced methodologies for accurate diagnosis, initial patient stratification, and follow-up.

The first activity will be to harmonize protocols and define standard operating procedures for clinical assessment, imaging, neurophysiology, the collection and storage of biological samples, and initial molecular profiling. A specific AD, PD, ALS database and a Technopole-dedicated biobank will be created for this. A parallel activity of the Outstation will be to re-evaluate the established AD, PD, and ALS cohorts and the relevant biological material in order to select samples that can be made available to the Human Technopole for advanced genomic studies. Based on its recruitment power and the already established cohorts of patients, the Outstation is expected to provide the Technopole with 2,000 biological samples from extensively characterized patients in the first 3 years of activity.

The "Carlo Besta" Neurological Institute has an internationally recognized neuropathology laboratory with extensive experience in neurodegenerative diseases. This facility will be made available to the Outstation for confirming diagnoses, staging the disease process, defining the disease subtype, and collecting brain material for morphological and molecular studies.

This network will be also crucial in establishing genomics-based multicenter clinical trials.

Brain Bank.

An important part of the project will involve the molecular analysis of post-mortem brains. We will thus promote and organize the collection of post-mortem brain tissues, setting up a national Brain Bank. This will benefit the wider Italian medical and research communities. Patients' data and their medical history will be included in the Human Technopole database together with the molecular and neuropathological characterization of samples. The project will include post-mortem brain tissues and patient data that are already available in the Hospital Network.

Genomics.

Here, we aim to study the genomic basis of neurodegenerative diseases by sequencing the genomes of large cohorts of patients and post-mortem tissues. This will be instrumental for a molecular-based classification of patients and in identifying drug target combinations for drug development. Short-length reads will be combined with single molecules, long-range sequencing to carry out de novo assemblies of genomes to identify structural variants. While genomes from peripheral tissues will be sequenced at traditional 30x coverage, neuronal genomes from post-mortem brains will be analyzed at higher depth. In five years, we expect to sequence the genome of 3000 patients and 3000 controls. These will be then analyzed by DCS and CLSC in the search for a genomics-based stratification of patients.

Biomarkers.

Neurodegenerative diseases have a long preclinical asymptomatic phase, wherein a silent biological process leads to progressive neuronal damage. The disease becomes clinically apparent only when it reaches a certain threshold, but it may be too late to halt it at this stage. Biomarkers play a pivotal role in the field of Personalized Medicine. Biomarkers are in vivo imaging or ex vivo molecular data that may be used for early/preclinical diagnosis, for prognostic scope, or to target and follow up on therapies. They may be disease-specific or indicate neurodegenerative mechanisms common to different disorders. Biomarkers may signal the presence of neuronal dysfunction or cell loss and indicate the severity of the disease.

By taking advantage of the cutting-edge knowledge within the Clinical Neuroscience Centers affiliated to the NGC, a list of biomarkers will be evaluated for each individual, using targeted expression analysis, in vivo imaging, and protein biochemistry. New markers will be sought with an extensive gene expression analysis performed on selected populations of blood lymphocytes and upon isolation of exosomes. In five years, we plan to carry out biomarker analysis, including blood transcriptome, of 1000 AD patients, 1000 PD patients, and 2000 healthy subjects.

Epigenetics.

Regulated chemical modifications of genes or gene-associated proteins play an important role in learning, memory, and behavior. Aberrant epigenetic mechanisms are involved in many neurological diseases, with recent data supporting epigenetic modulation in neurodegeneration. Epigenetic hallmarks will be analyzed in patients' peripheral tissues, iPSC-derived neurons, and post-mortem brains. This information will be included in the Human Technopole database and used to classify patients, study the molecular mechanisms of disease, and develop epigenetic-based therapeutics.

Microbiome.

Brain aging is affected by several biological mechanisms that are influenced by symbiotic microorganisms. Importantly, the microbiome is a crucial determinant of host immune system function and homeostasis. Here, we will investigate the relationship between diet, the microbiota, and neurodegenerative diseases. In this context, a genomic analysis of patients' microbiota will be carried out in collaboration with AFNGC.

Nutrigenomics.

In collaboration with AFNGC, we will analyze the relationship between genomic sequence variants, epigenetic marks, and diet to elucidate its contribution to neurodegenerative diseases.

Bioinformatics.

Pipelines will be optimized to assemble short and long reads, for single-cell genomics, and to identify the genome's somatic structural variants. We will analyze differential gene expression in peripheral tissues for a variety of RNA classes, including lncRNAs and repetitive elements. Special emphasis will be devoted to identifying differential Transcriptional Start Sites usage and gene networks.

RL2. Models & Mechanisms.

Vision. We will identify selected classes of patients with a specific repertory of genomic mutations. We will then use genome editing technologies to create human iPSCs and multigene iPSCs and animal models for multiscale studies of these selected classes.

iPSCs are becoming state-of-the-art tools for studying the neurons of patients with neurodegenerative diseases. Generated from adult cells, they can be differentiated in a reproducible manner into specific neuronal cell types, which maintain the genomic repertory of the individual of origin. iPSCs can then be used for in vitro pharmacology, molecular analysis, and electrophysiological characterization, including with multi-electrode recording devices. Importantly, they can be modified by genome editing of their disease-associated genomic variants in order to study the relationship of each of these sequences with cellular phenotypes.

Multigene animal models will be created with CRISPR/Cas9 technology and studied at different levels of organization from molecules to behavior. These models will be used to unveil molecular mechanisms of disease and to test new therapeutics in vivo.

iPSC Lab.

The iPSC Laboratory for Neurodegenerative Diseases will be responsible for the morphological and functional analysis of iPSC-derived neurons from patients recruited in *RL1*. The goal will be to identify disease-specific deficits for correlation with gene expression changes and identified genome variants. The Lab will establish CRISPR genome editing to correct disease-specific mutations, accelerate gene targeting, and develop reporter-cell lines. It will work in close collaboration with the F4's iPSC Platform. It will develop strategies for regenerative therapies in vivo.

Mesoscale Lab.

A major challenge for modern neuroscience is to develop methods to integrate data across scales and datasets. To this end, an interdisciplinary platform must be assembled to integrate infrastructures for viral tracing, optogenetics, two-photon microscopy, and freely moving analysis in rodents that allows direct correlations between neural activity and behavior. By perturbing and recording neural activity on a fine scale in CRISPR/Cas9-edited animal models, we will be able to assess changes in the excitability, connectivity, and complexity of cortical and subcortical networks (the basic parameters of brain function). This activity will be integrated with an advanced macroscale human electrophysiology platform, already available at the University of Milano. This platform uses magnetic perturbations and high-density electrophysiological recordings to study changes in human cortical excitability, connectivity, and complexity in both health and disease. This will offer a straightforward opportunity to translate basic animal research into activity involving patients directly.

Cognition and Plasticity.

The synaptic platform for neurodegenerative diseases will be a state-of-the-art neurophysiology laboratory specialized in studying synapse formation, function, and plasticity. It will investigate the role of novel disease-associated gene variants in synaptic function.

Neuroinflammation.

While neuroinflammation is a key element for brain homeostasis, it also plays a crucial role in neurodegenerative diseases. It is therefore important to study genomic variants affecting genes involved in inflammation, to detect the presence of inflammation in multigene models of disease, and to target it with new therapeutic interventions.

RL3. Translational Neuro Genomics.

Vision. Genomics-based diagnostics will be combined with new pharmacological treatments for neurodegenerative diseases. NGC will define a map of gene signatures of drug activity on neurons. It will focus on two goals: delivering nucleic acids to the nervous system for RNA-based therapy, and using genome editing technologies for brain repair. Building on the experience in the Brain Initiative and the Human Brain Project, it will develop innovative technologies to map human circuit function in vivo and set up new Brain-Machine Interfaces.

RL3 will be carried out in close collaboration with CLSC. Importantly, RL3 will be in close contact with the Clinical Neuroscience Centers in RL1, who will be setting up multicenter clinical trials for genomic-tailored drugs.

Clinical Genomics.

This activity will focus on translating information obtained in RL1 into tools and protocols that can be reproducibly applied to a large number of individuals for the Precision Medicine of neurodegenerative diseases. Innovative devices and software will be set up in order to i) calculate the risk of developing a neurodegenerative disease; ii) diagnose a neurodegenerative disease; iii) monitor disease progression; iv) predict drug response, and v) stratify patients. These new tools will have great potential for entering the market and changing the landscape of Precision Medicine in neurodegenerative diseases.

Mouse Clinic.

Genetically edited mouse models of selected classes of patients with neurodegenerative diseases will be used to analyze the effects of potential therapeutics in vivo. Phenotypes will be scored with behavioral tests and molecular analysis.

NeuroDrugMap.

In the Drug Connectivity Map, drugs are connected in a network if they elicit a similar transcriptional response. The map comprises more than one thousand compounds that can be subdivided into drug communities, i.e. groups of very similar drugs with a similar mode of action, but which are very different to other drugs in the network. This approach has been instrumental in repositioning drugs for new diseases and in identifying drugs whose transcriptional response is anti-similar to gene signatures in human diseases. Unfortunately, gene expression data for neuroactive drugs seem to be transcriptional noise, since drug targets are not expressed in the cell types used for experimental gene expression analysis. Here, the transcriptional response to FDA-approved drugs and natural compounds will be studied in preparations of primary mouse cells and selected neuron types differentiated from human iPSCs. This will provide the first NeuroDrug Connectivity Map for drug repositioning and for comparison with gene signatures from selected classes of patients with neurodegenerative diseases.

Nucleic Acids Therapy.

The post-genomic era has revealed the existence of a large repertory of noncoding RNAs and repetitive elements that play a fundamental role in cellular homeostasis and dysfunction. These may offer unprecedented opportunities to modify gene expression at the right time in the right space in vivo, providing an almost unlimited reservoir of new potential pharmacological agents. By hijacking their modes of action, the druggable genome can be extended to include regulatory RNAs and DNA elements in a scalable fashion. Beneficial effects can be obtained by inhibiting and/or increasing gene expression, depending on disease type and drug target. Here, we will focus on designing a new generation of nucleic-acid-based drugs, which are tailored to the mutations found in neurodegenerative diseases.

Genome Editing.

The recently developed CRISPR/Cas9 system of RNA-guided genome editing is revolutionizing genetics. Unprecedented opportunities to modify genomes have been created by the ability to recognize virtually any sequence in the genome and to introduce a controlled break. In this context, the ability to repair a gene mutation or delete a pathologically expanded sequence could, in principle, be curative. Given the current knowledge, several technical aspects must be addressed to efficiently and precisely modify the genome of post-mitotic cells such as neurons. This is a prerequisite for applying this technology as a therapeutic strategy. Here, we will focus on optimizing genome-editing technologies for neurodegenerative diseases.

Neurodelivery.

This activity will focus on optimizing delivery of nucleic acids into selected neuronal cell types involved in neurodegenerative diseases. Currently, the application of RNA therapy to neurodegenerative diseases is seriously hampered by difficulties in delivering to the correct site within the brain. We will assemble a coordinated effort in engineered nanosystems capable of carrying nucleic acids, delivering them to neuronal cells, and modifying endogenous gene expression levels for therapy.

Brain Technologies.

Brain-machine interfaces arise from the ability to record the activity of neural circuits operating in vivo and to use them to operate artificial devices in order to restore brain functions in patients suffering from invalidating neurological conditions. Functional data in humans are mostly recorded with noninvasive techniques, such as functional magnetic resonance imaging, electroencephalography, or magnetoencephalography. These signals cannot resolve individual cells or cell types, which precludes understanding of the fine cellular mechanisms that generate brain dysfunction. In turn, a key issue for brain-machine interfaces is how neuronal population activity determines behavioral decisions. However, the role of these spatial and temporal patterns in human disease is still largely unknown. Here, we plan to exploit innovative technologies (such as new optogenetic probes, diagostics nanosystems, ultra-high-density C-Mos micro-electrode arrays) to: i) predict the activity of specific cell classes from noninvasive human recordings; ii) dissect human brain function with unprecedented spatial and temporal resolution; and iii) understand the neural code. This information will then be used to develop a new generation of neuroprostheses to trasmit multiple neuronal ensemble signals to brain-machine interfaces that can restore abilities in patients with neurodegenerative diseases.

F

C3: AGRI FOOD AND NUTRITION GENOMICS CENTER (AFNG)

Vision: Health promotion and disease prevention is a social and economic priority as we face increasingly strained healthcare systems, an aging population, and the high individual and economic costs of diseases. Better nutrition appears to be the main effective, cost-efficient prevention strategy, particularly for chronic noncommunicable diseases (NCDs), such as diabetes, cardiovascular diseases, or cancer, where unhealthy diet is recognized as one of the major risk factors. The whole food chain is involved in the provision and consumption of healthy diets, and is interlinked with many other areas such as healthcare, the economy, the environment, and individual lifestyles.

Food production is the main outcome of agricultural activities. However, the current forms of both intensive and extensive agriculture are intrinsically fragile in terms of sustainability. Solutions based on high levels of scientific knowledge are needed to develop agricultural systems that require less energy input. The challenge is complex. On the one hand, food research and its targets can be designed with a reasonable probability of being correct. On the other hand, these models must be adapted to research priorities that consider their integration into sustainable agricultural systems. In other words, a food program must consider the centrality of sustainability to agriculture. This approach should also consider the changing socioeconomic conditions and their impact on the ecology of agrosystems, in particular when adopting modern systems biology approaches.

The beginning of the new millennium has coincided with a technological revolution in all fields of biology: the genomics era, based on high-throughput technologies, is revealing the metabolic networks that shape plant, animal, and microbial traits of agricultural interest. In particular, a massive amount of data on the genetic determinants of useful organisms is being produced, in quantities that are unprecedented for biology but essential for both basic and applied studies. There has been a parallel expansion in the amount of data concerning the impact of nutrition on human health. This has created a web-based resource that could potentially integrate omics data for the design and production of the foods of tomorrow. We now need to be able to understand, interpret, and use the available and future data. This can be achieved by developing systems-based methodologies that fully enable cross-disciplinary integrative approaches. Specifically, big data management technologies must be integrated with an increased understanding of the interactions between the agroecosystem drivers that are focused on food production. By achieving this, we seek to lead the strategic knowledge-based development of novel healthier foods derived from self-sustainable agricultural systems. In terms of the quality of its agricultural produce, its food products, and its diet, Italy is a well-established world leader. The country is thus well poised to play a leading role in providing the needed solutions. This novel knowledge hub will be instrumental in addressing the national and European challenges of ensuring a safer, healthier, and more sustainable food production for human and animal nutrition.

To achieve these goals, we will invest in the generation, integration, and interpretation of data. Supported by predictive models and systems biology approaches, we will transform data into an understanding of the physiological and molecular impact of nutrition on human health. This will allow us to develop functional solutions for disease prevention and personalized nutrition with a scientifically sound basis. Together with advances in ICT, nanotechnologies, remote sensing and robotics, the genomics of the living components of agricultural systems can introduce transformational innovations for a more sustainable food system. Combining these approaches will produce a "farm to health" concept with safe, sufficient, affordable, and healthy dietary components. Our goal is the technology-driven rational design of bioactive foods and the development of self-sustainable nutritious crops that reduce agriculture's need for chemical and energy inputs. We will explore diet-health homeostasis by considering how the food metagenome and immune function interact, including the poorly explored analyses of the mycobiome and the role of fermentation-derived microorganisms in health. We will integrate with the other HT centers and cross-fertilize our research with systems nutritional metagenomics, metabolomics, and clinical studies in order to define the novel landscapes of Precision Medicine and address the self-sustainability of health in this era.

Scientific Structure: The AFNG Center will develop 4 Research Lines*: Crop Genomics and Biotechnology; Functional and Systems Metagenomics and Metabolomics; Food, Host, and Microbiota Interactions*; and

Personalized Nutrition. In addition to the "Common Technology-Development Platforms" described in Facility F1, the Center will implement the following enabling technological platforms specific to its activities: *i)* greenhouses and high-throughput plant phenotyping; *ii)* plant biotech lab and genome editing; *iii)* plant cell culture; *iv)* plant genome analysis (PacBio) and plant cytogenetics (BioNano); and *vii)* microbial biotechnology.

The AFNGC will comprise a central research facility at the EXPO site with Joint Laboratories and Outstations. These will include the University of Milano (*Smart Crops and Smart Food*), FEM (*Genomics and Post-genomics of Fruit Crops*), CREA (*Seed Genomics, Precision Agriculture),* IIT *(Precision Agriculture* and *Innovative Packaging*), and PTP (*Bioinformatics* and *Crop Biotechnology*). It will create synergies with CSMD to develop technological innovations and translational activities to improve food quality and safety.

Research lines and activities:

RL1. Crop Genomics and Biotechnology.

Vision. To exploit new traits of potential agricultural and nutritional interest, the Center will adopt an integrated approach including NGS resequencing, Mega-QTL maps, genome-based in silico gene discovery procedures (validated via reverse genetics approaches), and advanced genetic transformation and editing. This will be integrated with a high-throughput plant cell regeneration platform to exploit the major opportunities offered by advanced knowledge of plant genomes. Activities will focus on: i) crop plants, particularly of fruit trees like grape and apple (for which a MAS (marker-based selection) system will be made available for quality traits), and fruit and seed development; ii) substitution lines of maize to associate specific chromosomal fragments with heterotic effects, followed by cloning of the putative genes supporting such effects, with reference to Specific Combining Ability; and *iii)* breaking breeding barriers by developing strains with wide sexual compatibility.

Seed Development and Quality.

Next-generation genomics will elucidate how critical developmental stages are regulated for fruits and seeds. Of particular interest are the seed endosperm of cereals, which produce 60% of all proteins and calories necessary for human nutrition. Molecular markers, haploblocks, and haplotype combinations have already been integrated, at least in part, into the selection process for new improved varieties. Exploiting our access to a large-scale facility for big data storage and analysis, we will use genomic selection algorithms and pedigree-based analyses to: i) use information across a range of crop species, moving from crop to crop when approaching a problem; and ii) use synthetic biology exercises to implement comparative genomics. We will develop a computational biology structure dedicated to plants, with priority given to *Solanaceae*, *Rosaceae*, and Mediterranean crops, which are of particular interest to Italian agriculture.

Self-supporting Crops.

Despite significant improvements in crop yield potential and quality in recent decades, the expected global climate changes raise concerns for food safety. In this changing environment, yield stability and productivity will require the development of *smart* varieties and cropping systems better adapted to water stress. To achieve this, breeding effort should focus on selecting high-yielding genotypes with enhanced water uptake from the soil, reduced water loss, and increased yield stability under drought. This approach requires the integration of omics technologies, quantitative genetics, genome editing, precision phenotyping, and biochemistry. To generate crop plants resistant or immune to major pathogens, we will use molecular knowledge of plant-pathogen interactions mediated by the genomic gene family NBS-LRR (R-genes), which encodes receptors for pathogen-produced signals. An additional avenue is to consider disease susceptibility S-genes. Knocking out these genes via gene editing leads to plant immunity against all strains of a given microbial parasite. An additional goal is to develop MAS protocols for S- and R-gene transfer for use in the practical breeding of superior varieties that require fewer agrochemical treatments.

Phytochemicals and Micronutrients in Novel Smart Foods.

We will consider the enhancement of nutritional value ("smart foods"), including enhanced micronutrient content, reduced anti-nutritional components, producing therapeutics for customized medicine, and plants with enhanced prebiotic capacity. Particular attention will be devoted to ncRNAs of plant origin, including their cross-kingdom signaling, health-promoting properties, modulation of inflammation, and healthy immune functions. We will also explore the complexity of the diet-health relationship by developing near-isogenic model foods that differ only in the type and quantity of specific micronutrients. These models will allow us to develop functional foods/nutraceuticals with a scientifically demonstrated benefit in a diet-related disease.

In-field and Post-harvest Technologies.

We will adopt modern computational agronomy based on biophysical modeling approaches to ensure that soil, crops, and livestock are managed in order to precisely control nutrients and pesticides, whilst enhancing support for agriculture, ecosystem resilience, and associated ecosystem services. To develop precision agriculture models describing more self-sustainable agricultural systems, we will integrate critical advances in sensor technology, intelligent actuators, remote and proximal sensing, computational capacity, and agro-biotechnology. We will develop systems to integrate, on a single platform, product traceability/quality data from the entire value chain. New biomaterials from bio products and agro-wastes will be considered for food packaging with advanced functional and quality characteristics.

RL2. Functional metagenomics and metabolomics.

Vision. The microbiome is the complex microbial community that symbiotically interacts with all living organisms. The metagenomics revolution could hold the potential to improve human and plant health via directed intervention in the human microbiome. We will use evolution to understand how molecular signals dynamically regulate the web of interactions that subtend the equilibrium of bacteria, viruses, fungi, and their hosts. Plant-borne and food-borne microbiota can have a drastic impact on the human microbiota and could be a tool for modulating the human microbiota. Strain-level global analysis of the human microbiota will allow us to construct strain-level metabolic maps and to trace the microbial flow from food and the environment to the human gut. We will model the gene and metabolite networks, which link the patterns and functions of microbial species to health and disease. This will facilitate the future design of personalized probiotic and phage-therapy-based interventions to selectively alter specific components of the microbial web. By deepening molecular understanding of the host specificity of biocontrol agents, we can reduce nontarget effects through indirect interactions and food-web subsidies.

Human Systems Metagenomics.

A catalogue of the human microbiota has been generated via throughput sequencing technology. With the microbiota, hundreds more genes must be included in our genome. The human microbiota has provided a new set of functions that our organism has learned to exploit over the years. This has enhanced human genetic variation between individuals. Changes in microbiota composition have been associated with several disorders, including cancer, metabolic diseases, and neurodegenerative diseases. We will consider microbiota from large cohorts of patients with different diseases, individuals from rural and industrial areas, twins, athletes, and vaccinated and nonvaccinated individuals. We will culture and identify, at the strain level, anaerobic bacterial species found in rural settings and in vegetarians. We will develop markers of pathogenicity at the levels of strain and species, revolutionizing food safety and pathogen detection approaches. Priority will be given to i) disease-associated microbiota from colorectal cancer patients (with and without food supplementation) and from patients with prostate and bladder cancers (in close collaboration with OGC); and ii) microbiota from AD, PD, and ALS patients (in collaboration with NGC). We will assess whether there are microbiota profile patterns associated with these pathologies and whether these can be altered with diet.

Evolutionary Systems Metagenomics.

A collection of fungal strains from rural regions of Italy and the Mediterranean will be generated, including commensal fungi from plant mycorrhizae, fermented foods, human gut, insects' gut, and opportunistic pathogens, focusing in particular on their immunomodulatory potential in plants and humans. From healthy individuals, we will generate reference values for the ratio between fungi and bacteria and between the different genera. Fungi will be isolated and fully characterized in order to generate useful references for next-generation probiotics for human and plant use. We intend to create an extensive centralized culture collection as a reference for developing strain-specific markers. We will achieve this using high-throughput culturomics (in particular of the largely untapped anaerobic and fungal component of the microbiome) in parallel with high-throughput phenotyping (omnilog). We will use bioinformatics methods for metagenome analyses in order to develop pathogenicity markers at the strain level. The results will facilitate a revolution in the methods currently used to discriminate pathogens from commensals. They will allow next-generation strategies to be developed for food safety assessment, increasing consumer care and maximizing food quality. Databases will be cross-annotated and integrated at the European and international level within the Elixir ESFRI. We will also implement the forward and reverse genetics toolboxes available for strains of the *S. cerevisiae* model. These genetics toolboxes will facilitate systems biology approaches to elucidating the host-microbiome interactome, providing novel opportunities for the fermented foods industry. These toolboxes will be made available on a collaborative basis.

Agri Food Metagenomics.

The plant microbiota contributes to plant fitness for environmental conditions and thus to the stability and productivity of the agricultural and natural ecosystems. Metagenomics and meta-transcriptomic approaches will be used to address host/pathogen and rhizosphere/root molecular interactions. We will also characterize the microbial populations and their functions as they relate to the biological quality of soils. This will allow us to identify bio-inocula and substrates that can make soil microbial populations more beneficial in term of bio-fertilization and bio-protection. This will also allow us to adopt appropriate soil bioremediation methods. Microbiota community structure data will be used to improve microrganisms' in vitro culture and artificial consortia for functionality. The data may also be used to identify new bioactive metabolites for industrial production. The final aim is to develop new bioactive compounds against plant and human pathogens. The compounds should be based on natural, biodegradable, and renewable molecules. Microbial-driven epigenetic modifications and their effects on plant metabolome and proteome will be investigated in relation to the production of defense components. This will allow us to assess the use of nonpathogenic components of the microbiota and mycobiota as plant vaccines.

Metabolic Networks.

Nowadays, high-throughput nontargeted analytical technologies are available to elucidate the complexity of plant cells' biochemical networks. The accumulating metabolomes and proteomes contain a wealth of information that can reveal the inner cogs of cell machinery and how plant cells interact with each other and with the environment. This has revealed metabolites and protein/peptides with key functions in determining plant fitness, including interacting with beneficial hosts (the plant microbiome) or pathogens. New mechanistic interactions can be identified at the molecular level by combining metabolome/genome interactions, metabolome/proteome resistance phenotypes, and epigenetic effects on plant metabolome and proteome. Metabolic networks will be investigated and refined to develop a functional genomics of key micronutrient biochemical networks and secondary metabolites. We will combine MGWAS approaches with smart screening for biological functions to identify novel phytochemicals via integrative nontargeted approaches. The information will be annotated and cross-referenced according to the most advanced international standards within the context of the JPI-ENPADASI project and the Nutritional Phenotype Database in Elixir.

RL3. Food, host, and microbiota interactions.

Vision. Human phenotypes associated with different microbial communities have been described as being strongly dependent on diet, with three diet-dependent enterotypes identified. The microbiota-food-host interaction is fundamental for the wellbeing of our organism because the microbiota contributes to several important functions. These functions include maturation of the immune system, protection against infectious agents, brain function and behavior, and detoxification of toxins and xenobiotics. Dysbiosis is a deregulation of the microbiota composition. It has been associated with several mucosal and systemic diseases, including metabolic disorders, neurodegenerative diseases, and cancer. However, it is not yet clear if dysbiosis has a causative role or is the consequence of the disease. In this research line, we will conduct functional studies on the interactions between microbiota, food, and host. These will include studies on immune modulation, nervous system activity, and their mutual relationship. Activities will involve joint programs and synergic collaboration with scientists involved in *Fundamental Genomics* at OGC and *Neuro Genomics* at NGC.

Dietary Modulation of Immunity.

Several foods or bacterial components have been described as having a strong effect on the immune system either by potentiating the immune response or by inhibiting inflammation. We will use different experimental settings, ranging from simple systems (isolated immune cells, including dendritic cells and T lymphocytes, epithelial cells, and the recently emerged stromal cells) to more complex systems such as ex vivo human organ cultures or animal models. Our aim is to identify products with different immune properties that can potentiate or dampen the immune response. These products could subsequently be used in diverse applications, including increased protection against infectious agents, improved vaccination, or reduced risks of autoimmunity, cancer, or metabolic syndrome.

The Gut/Liver/Brain Axis.

Within the overarching goal of self-sustainability of health, one of the main objectives is to understand how and if inflammation (or the modulation of inflammatory responses) can begin in the gut and spread systemically to other organs, such as the liver or the brain (gut-liver-brain axis). Recent studies supporting this theory have highlighted a possible correlation between the microbiota and several immune disorders. Defects in gut barrier permeability have been associated with increased liver damage, metabolic syndrome, and autism spectrum disorders. This suggests that studies in this direction will shed light on gut-derived systemic disorders.

Nutritional Epigenomics.

A growing body of evidence suggests that food components and the microbiota can impact the epigenome of individuals. Nutrients can affect DNA methylation and induce histone modifications, impacting gene expression. We will study the effect of food, the microbiota, and their metabolites on the human epigenome in adults and in newborns following 'controlled' maternal nutrition. It is important to understand if exposure to certain nutrients, which play important roles in histone modifications and methylation, can impact the fetus through maternal transmission. Our goal is to identify patterns of epigenetic modifications associated with certain categories of foods or microbes, which could be transmitted vertically or horizontally between individuals of the same family (e.g. siblings who live separately or in the same environment). This activity will be carried out in close collaboration with NGC.

Noncoding RNAs (ncRNAs) and Their Role in Diet/Health Relations.

Here, we will focus on the function, bioactivity, and biodiversity of ncRNAs. In particular, we will study their cross-kingdom signaling, health-promoting properties, and ability to modulate inflammation and healthy immune functions. These studies will be integrated with investigations into the protective anti-cancer function that bioactive complexes in food exert by modulating ncRNA expression. This line of investigation will elucidate the mechanism of action of these biomolecule complexes, focusing on the effects mediated by microbiome modifications. Moreover, we will target ncRNAs for pathogens in food as a mechanism of pathogenicity and to produce efficient biomarkers for diagnostic tools. A parallel research line will apply genome-editing technologies to improve the health-promoting properties of cultivable plant varieties and beneficial microbes.

Functional Foods and Nutraceuticals.

One challenging goal is to design novel enriched functional foods with optimal bioavailability of nutrients and health-promoting effects, particularly by combining prebiotics and probiotics. Plants, fruits, and vegetables will be screened for novel molecules with health-promoting properties, focusing on anti-inflammatory molecules and molecules improving immune function. We will analyze the role of plant processing, conservation, and food preparation procedures in order to optimize the levels of health-promoting bioactive molecules present in transformed products and foods. We will also study the role of environmental, plant-associated, and foodborne microorganisms during food preparation (fermentation), performing an extensive integration of the microbial and food metabolomes.

RL4. Personalized nutrition.

Vision. To achieve the self-sustainability of health, we must investigate the interplay between the human-genome-encoded determinants of immune function, diet, and the metagenome. Next-generation nutrition will identify personalized functional foods combined with personalized probiotic therapies to be administered to healthy individuals with certain metabolic patterns. This will allow diet to be tailored to each individual to achieve health benefits. We will set up parallel massive human genome and metagenome sequencing. This will be associated with the precise assessment of food and body fluid metabolome composition in longitudinal and intervention studies to clarify and manipulate the relationship between food, microbiota, and human health. The results will be used to identify markers of disease predisposition in unbiased approaches. We also propose to identify a formulation of Mediterranean-diet-based microbial strains for use in fecal transplantation studies. This formulation will be based on a cocktail of bacterial and fungal strains, combined with specific fibers and phytochemicals active as prebiotics, which can reconstitute a healthy microbiota. This formulation could be administered to patients with different dysbiosis-associated diseases, such as inflammatory bowel disease. We will thus conduct prevention and intervention studies on large cohorts of patients, using personalized food-microbiota intervention strategies. Activities within this RL will involve joint programs and synergic collaboration with scientists involved with *Therapeutics* at the Onco Genomics Center and *Translational* and Neuro Genomics Center.

Support-care Nutrition.

Elderly individuals, pregnant women, newborns, athletes, patients undergoing treatment – all need 'special' nutrition that is tailored to their health and immune status. Special emphasis will be given to designing and developing tailored nutrition to target these defined populations (the elderly, pregnant women, athletes) with functional foods, such as fermented foods, probiotic supplements, and symbiotic supplements. The aim is to potentiate or dampen the immune response, using support-care nutrition to affect the individual microbiota and modify the metabolome according to the target population.

Nutritional and Clinical Metabolomics.

Dietary components are a major determinant of protection or susceptibility to disease. This general concept applies to diverse human pathologies, including cardiovascular disease, neurodegeneration, and cancer. This research line's general objective will be to integrate the investigation of dietary components with an analysis of their impact on metabolism using the omics approaches available at HT. Areas of integration will include investigating the impact of diet on the microbiome, microbiota metabolism, host metabolism, and host defense. A major thrust of this research line will be to integrate state-of-the-art technology with large-scale epidemiology, taking advantage of the unique range of cohorts available from large population studies in Italy.

Preventive and Intervention Nutrition.

Identifying risk populations using microbiota and nutrition lifestyles is a recently proposed and novel concept. Our aim here is to monitor healthy individuals and individuals at risk of developing cardiovascular or neurodegenerative disorders in terms of their health, nutritional, and exercise habits. This will be achieved by developing devices to evaluate the type of food, number of calories, metabolic output, physical exercise, and quality of sleep of each individual, and correlating this with a long-term follow-up. By identifying nutrition and lifestyle patterns that correlate with predisposition to disease development, we can design preventive and/or intervention nutritional studies.

Microbiota Intervention.

Dysbiosis is a process found in several disorders, including immune-related, neurodegenerative, and metabolic disorders. To prevent or treat dysbiosis-related disorders, fecal microbiota transplantation (FMT) is carried out to reconstitute a 'healthy' microbiota. FMT has been successfully used to treat antibiotic-resistant *Clostridium difficile* infections with highly promising results. The aim of this research line is to standardize FMT. We will begin with healthy donors to identify the characteristics of a healthy microbiota. We will then seek to identify a cocktail of microbial species that can reconstitute a healthy microbiota. This cocktail would serve as a universal microbiota for transplantation to patients with different dysbiosis-associated diseases. These activities will require the storage of biological samples (including blood, feces, and urine), a registry of donors, and a repository for the isolated microorganisms. We will focus on microbiota associated with traditional fermented foods. These will be bio-banked and preserved because they are a potentially rich source of live probiotic microorganisms of environmental origin.

C4: DATA SCIENCE CENTER (DSC)

Vision: The mission of the Data Science Center is to develop a scientific framework to efficiently extract information from large sets of heterogeneous data (both structured and unstructured, possibly noisy) , with the aim of generating structured knowledge. Correlation patterns hidden deep in datasets are the key to transforming data into intelligence so it can be used to make predictions about the system in question. A virtual landscape of the system is thus generated, which serves as an efficient toolkit for decision makers and policy makers seeking to intervene. Within the Human Technopole project, this strategy and methodology will seek to address the profound questions posed by Precision Medicine. The DSC will be headquartered at the Expo site with an Outstation at the ISI Foundation. It will cooperate with the Università Statale di Milano, Università Bicocca, and Politecnico di Milano.

Scientific Structure: The Data Science Center will follow four main research lines:
RL1 Basic Issues in Data Science: 1) data management: acquisition, representation, integration, visualization; 2) information modeling: semantic, context-based enhancement; 3) information retrieval: data mining (DM), knowledge discovery; 4) artificial intelligence (AI) methods and machine learning (ML); 5) topological data analysis (TDA); 6) data trustworthiness, predictions; and 7) knowledge and context-driven data analytics, uncertainty modeling in data and processes.

These methods will be applied to:

RL2 Data-driven Genomics: data science methods for the analysis, model construction, simulation, and structural prediction of genomic data. The goal will be to understand genome variations in relation to clinical data and how they can be used in Personalized Medicine for cancer and neurodegenerative diseases.

RL3 Data-driven Neuroscience: topological data-handling methods for analyzing brain structure, functional correlations, and dynamics; and models of brain dysfunctions.

RL4 Digital Epidemiology: dynamics of health and disease distribution/diffusion in human populations; correlation of disease pathways and environment (nutrition network, human contacts); and monitoring platforms.

RL5 Ethical Issues in Data Science: privacy, (cyber) security, data anonymization, and the ethical challenges of Precision Medicine.

RL2 to RL4 closely connect DSC to all other Centers in an ongoing and mutual exchange of data and information. In particular, DSC will work closely with OGC, NGC, and AFNGC to analyze genomic and patient data. DSC will rely on CLSC for all aspects connected to algorithm implementation, software optimization, and high-performance computing, and it will rely on CADS for analyzing and interpreting social impact data. RL5 is transversal with respect to all Centers.

Research Lines and Activities:

RL1 Basic Issues in Data Science.

Vision. Big data (BD) impose a strenuous data management challenge, stretching the bounds of modern computer science (new computing paradigms beyond Turing machine; interactive computing; new approaches to data mining, machine learning, topological data analysis) and modern complexity science (network theory, nonlinear causal inference). Conventional approaches to complex problems in basic science (in particular the life sciences and social sciences) must pay more attention to basic conceptual and universal BD features.

ML is the branch of AI aimed at constructing and studying systems that can learn from data. It has two main goals: i) representing data instances and functions evaluated for these instances, in order to build machine-learning capacity; ii) extending this capacity to perform efficiently for unseen data instances. Both goals pose profound conceptual questions, since they define 'learning' in operational rather than cognitive terms, and raise such fundamental issues as the concept of 'thinking machines'.

ML focuses mostly on prediction based on properties learned from the training data. DM is the analytical discovery of unknown properties of datasets. The two overlap in many ways. DM uses a variety of ML methods, while ML uses DM methods for "unsupervised learning" or as pre-processing steps to improve learner accuracy. ML's ultimate goal is to eliminate the need for human intervention in data analysis. This is seldom achieved since the system's designer must specify how data should be represented and what mechanisms must be used to search for a valuable data characterization. Hence, the desired algorithm outcome depends on the type of input available when training the machine. ML algorithms and their performance are the challenge of theoretical computer science and computational learning theory. The training sets are obviously finite and much smaller than the dataset to be explored, so that learning theory is typically subjected to severe probabilistic bounds on the performance of its algorithms.

In 'Topological Data Analysis' (TDA), in contrast, appropriate mathematical tools applied to the 'space of data' seek to incorporate data into a topological setting, which then allows us to identify and control the hidden information patterns much more effectively. TDA is a theoretical framework allowing for the efficient exploration of large quantities of data by inferring information from global rather than local data space properties. It stems from integrating the profound mathematical aspects of topological analysis of the data space with formal language theory, theoretical computer science, and statistical physics.

TDA represents a profound evolution of conventional DM methods, whereas ML allows for the discovery of unknown features concealed by data, training the machine to recognize bits and pieces of the hidden information. "Unsupervised learning" may improve learner success in ML; however, it still requires some amount of a priori knowledge of what one is seeking. TDA requires much less a priori information on the system, but it is harder to implement. Both share the goal of optimizing the role played by IT in the following process:

data → information → knowledge → wisdom.

This is typically difficult because the set of all possible behaviors given all possible inputs is too large to be covered by the set of known samples. Furthermore, it is very hard to recover the hidden relations within the data space to recognize complex patterns. By extracting interesting patterns from the data, certain features can be learned. However, futher simulation effort is needed before intelligent decisions or predictions can be made based on these known features. Representing data in a topological framework enables us to explore the whole data space globally. We can thus better control its structure and the hidden information it encodes, as well as extracting the correct questions instead of answers to potentially incorrect questions. In addition, resorting to topological methods furnishes an efficient way of distinguishing noise from signal. The emerging theoretical framework should also provide innovative data-mining methods (based on a nonlinear field theory rather than on ML), which permit data spaces to be mined while reducing the a priori system knowledge required.

DSC will pursue research to extend data science in those research dimensions most suited to facing the problems and challenges of the applications at stake in RL2 to RL4.

Data science applications range from 'hard' sciences (genomics, neuroscience) to societal issues (health system control, digital epidemiology, food and nutrition) to the data of sensor networks. These includeall declinations of genomics and their cross-correlations, the brain and its dysfunctions, and the full cancer pathway dependence from environment to nutrition.

RL2 Data-based Genomics.

Vision. Ever since Crick and Watson's discovery, scientists believe that DNA contains a code. The code-breaking effort seeks to reveal the relationship between DNA, RNA, protein structures, their regulation, and phenotypes. This biochemical puzzle has been reduced to an abstract problem in symbol manipulation, since all molecular complexities are maps between messages in different alphabets and languages. Codes come from combinatorics, but their solutions belong to information theory. But when code theorists learn the language of genes, their techniques are all framed by a local vision of the problem. However, a major part of the relevant message is encoded in global features that need to be decrypted.

While inert matter essentially evolves a collective disorder, living matter evolves hierarchies of collective order thanks to simple cooperative functions. What is qualitatively exceptional about living matter is not only its complexity, but also the evolution of the global constraints governing those simple cooperative functions. The key lies in those processes whereby molecules function symbolically as records, codes, and signals, becoming a message. Communication between molecules differs from the normal physical interactions between molecules, which, according to physics, account for all their states of motion. A molecule does not become a message simply because of its particular shape, structure, or behavior; it does so in the context of a larger system of physical constraints, which can be thought of as a "language". We need to work at a different level to that of physics, focusing not on molecular structures but rather on the structure of the language through which molecules mutually communicate.

The central dogma of molecular biology states that information flows from DNA to RNA and ultimately to proteins: a single pathway of information is responsible for all cell phenotypes. According to this dogma, proteome complexity is how cells and tissues are structured with different shapes and functions. However, data released by large genomics projects have enormously increased the description of the (epi)-genome and its properties (mutations, peaks of expressions for protein binding, gene regulation, structural DNA properties). These data have challenged the classical view of gene expression regulation. Pervasive transcription gives rise to a large repertory of noncoding transcripts, whose expression is tightly regulated in space and time, including long noncoding RNAs (lncRNAs), small noncoding RNAs, and transcripts derived from repeats such as Transposable Elements (TE). These were once believed to have no function; but, on the contrary, they are crucial. The transcript of mammalian cells consists mostly of lncRNAs and repetitive elements that play a primary role in cellular homeostasis and dysfunction. While there are approximately 25,000 protein-encoding genes, more than 90,000 annotated human lncRNAs have been described. A TE is a DNA sequence that may change its position within the genome or duplicate itself at a different locus, acting as an endogenous retrovirus. Thus, as the gene expression process is progressively broken down, several questions emerge in the search for a direct association between DNA, RNA, and biological functions. These questions are: i) how do common DNA sequences relate to the results of transcription? ii) how does the correlation between transcription and translation take place? iii) what associates protein expression and nonprotein coding genes? iv) what is the role of the genomic structural variants? v) what is the influence of epigenomic patterns on phenotypes and their likelihood; what are the risk factors?

Functional data relating sequence to structure data (RNA secondary structure and protein conformation) exist at the (epi)-genomic and transcriptomic level and in medical-pharmaceutical databases. However, a functional analysis does not yet exist. To achieve this, we must relate sequence data to structure data in order to extract patterns between the epigenomic, genomic, and transcriptomic levels, i.e. RNA and protein functions.

Cells become cancerous when genetic and epigenetic reprogramming of complex regulatory networks bring about their immortality and uncontrolled division. Hence, cancer can be characterized by a specific tendency to maximize entropy by tumor cells. This is in contrast to normal cells whose main dynamical objective is homeostasis. To achieve their goals, cancer cells eliminate cell-cycle checkpoints and simplify their program. By analyzing cancer hallmarks from the variety of DNA mutations, we aim to find an entropic organizing principle at all levels of transformation. We will rigorously check this hypothesis using all types of data available, leveraging signal discovery in genes and in lncRNAs, and developing BD algorithms for cancer, including cancer progression models. This should allow practical conclusions to be drawn when narrowing down therapeutic interventions to those that can slow down or reverse cancer cell proliferation. Efficient data analysis algorithms can identify drug-actionable genomic targets, helping to develop innovative strategies for better anti-cancer drugs. Combining biomolecular knowledge and genomic data will allow chemotherapy drugs to be optimized for specific types of cancer. The end goal is a personalized therapy, which affects essential pathways with mechanisms that result in cancer cell death and that carry minimal side effects for normal cells. This

F

therapy must overcome intrinsic and acquired resistance, tumor heterogeneity, adaptation, and the genetic stability of cancer cells.

Using innovative data management and ML methods that can tackle the volume of data available, it will be computationally feasible to design therapeutic regimes for individual patients using most of the current chemotherapeutic drugs, whose molecular mechanisms are well understood. The ensuing drug optimization strategy will involve the identification of specific protein targets from interactome networks as well as specific algorithms and validation methods. This strategy should enable the use of a computational molecular-scale network approach to minimize the size and cost of clinical trials as well as the failure rates of potential cancer inhibitors.

Similar challenges must be confronted in the field of neurodegenerative diseases. The molecular bases of sporadic AD, PD, and ALS are still unknown and there are still no effective therapies. Their extreme heterogeneity is a critical roadblock, which limits our ability to apply Precision Medicine paradigms to these diseases. It is therefore crucial to identify molecular classifiers to stratify patients. Within this context, the Human Technopole offers a unique opportunity for data analysis to search clinical and genomics data on an unprecedented scale. We thus aim to identify patient groups according to combinations of gene variants, and to define molecular signatures as biomarkers for disease diagnosis and progression.

RL3 Data-driven Neuroscience.

Vision. The study of human brain networks using in vivo neuroimaging has given rise to the field of connectomics, which has been furthered by advances in our understanding of brain topology and function. The Connectome is the complete map of the brain's neural elements and their structural interactions, allowing for the complex integration and segregation of all the relevant information. Whole-brain data-based computational models can make inroads in understanding the link between structural and functional brain connectivity. Connectomics focuses on the disruption in neurological disorders and on how models can help to generate and predict the interactions and outcomes of brain networks across several timescales. Brain anatomy is the basic 3D framework, within which most types of neuroscience data provide information about organized local and long-distance connections. The paradigm of connectomics includes distinct subdomains for long-range pathways, which link patches of gray matter and generate the complete connectivity map within a single patch. Noninvasive neuroimaging charting of human-brain neural circuits, in parallel with full-genome sequencing, will enable hypernetwork DM to reveal detailed connectivity differences according to behavioral phenotype.

Although mental dysfunctions are heritable, they appear to be genetically complex and heterogeneous, and most of the genes that impart vulnerability have yet to be discovered. By efficiently navigating across the multiple tiers of data, we aim to capture the critical changes in neuron and/or glial functionalities that reveal ways to discover/prevent disease, finding common threads to all mental dysfunction symptoms. The proof of concept will be in groundbreaking models designed to use algebraic topology methods to study pathological brain activity based on large sets of fMRI data.

The goal of neuroscience is to understand the mechanisms underlying the boundless range of brain functions and to uncover the causes of a vast array of brain disorders. We need to develop methods to capitalize on the volume and diversity of data generated in this context in recent decades. Getting to grips with "neural choreography" (NC) will be central to this task. NC is the integrated collective functioning of neurons in brain circuits, their spatial organization, local and long-range connections, temporal orchestration, and dynamic dependence on interactions with glial cell partners. NC is a fully fledged and complex system that cannot be understood via a purely reductionist approach. Rather, understanding NC requires the convergent use of a variety of tools to gather, analyze, and mine information from each level of analysis and to capture the emergence of new layers of function (or dysfunction). The goal is to elucidate brain-function flow through the multiple layers of hierarchical organization, which operate at different spatial and temporal scales, and the alterations of this organization in pathological disorders. TDA-inspired computational models allow us to represent and predict the dynamic interactions and behavioral outcomes of brain circuits across a variety of timescales.

This places network science among the set of methodologies needed to characterize the brain's dynamics as typical of complex systems. To date, much of this research has been descriptive. Now, however, we have at our disposal the necessary conceptual tools for data-based brain computational science to make inroads in understanding the link between structural and functional brain connectivity and their potential breakdown in disease. The topological approach will enhance the characterization of key features of brain networks (seen as a simplicial complex) by allowing the description of coordinated multi-area activation patterns.

A new paradigm in brain sciences will be kick-started by a symbiosis of concepts coming from complex systems science, topological computation, AI, and the availability of large sets of neuroimaging data. This paradigm will be based on: i) collecting and fusing the brain's functional, genetic, and connectomic data across spatio-temporal scales, from the microscale of neuronal cultures to the macroscale of functional imaging; ii) developing theoretical tools to detect/analyze brain activity patterns, capturing functional and mesoscale distributed coordination patterns; and iii) developing synthetic data-driven whole-brain computational models that can generate personalized quantitative descriptions of individual brain activity.

By fusing functional imaging techniques and connectivity data, we will produce insights into the emergence of resting-state dynamics and the role of noisy dynamics and neuronal plasticity. Breakdowns in brain function can be mapped to disruptions in the connectivity structure; yet it is hard to identify mesoscopic biomarkers that can detect and discriminate brain dysfunctions and degenerative diseases at early stages and to follow their evolution in time. Novel techniques of data analysis and data fusion are necessary to produce a computational framework that can: i) integrate different data sources; ii) analyze them across different modalities (multilayer) at various levels of resolution (multiscale); and iii) integrate this framework into a system that can discriminate brain dysfunctions as they emerge.

Detailed data-based large-scale thalamo-cortical models have already proven their ability to span different anatomical scales: macro (white matter; global thalamo-cortical anatomy), meso (branching dendritic structures), and micro (neuronal circuitry; multiple thalamic nuclei). Very large numbers of neurons and synapses must be represented in a reliable form that can account for receptor kinetics, short-term plasticity, and long-term dendritic spike-timing-dependent synaptic plasticity.

Reliable, anatomically detailed, and consistent models of the cerebral cortex would facilitate a data-science-based approach to linking brain dynamics (and the underlying neuronal mechanisms) to perception and cognition. This would allow us to tackle the emergence of functions including micro- and meso-circuitry in local dynamics, synergy between multiple cortical regions, cortico-cortical connections, and synaptic plasticity. The macro objectives of RL3 are then: i) improved morphological modeling of complex brain networks, resorting to advanced mathematical tools to describe the system at different morphological scales; ii) effective coding and representation in terms of automata capable of recognizing the appropriate languages; and iii) multiscale growth dynamics and self-organization of flexible inter-areal functional connectivity and information routing.

A basic tenet in cognitive science is that the brain represents events in the external world by generating a virtual reality in *n*-dimensional state spaces associated with neuronal activity and by performing mappings between these state spaces. This complex, dynamic, and self-organized representation accounts for the response to an unpredictable environment. It must be broadcast to the networks responsible for further information-processing steps when information is selectively routed across other mesoscale or macroscale brain circuit (hyper)networks. How does the static inter-areal connectivity give rise to an information routing and a functional connectivity that are flexible and directed? A functional geometry of the cerebellum is needed. It should be based on all possible maps between the *n*-dimensional virtual state spaces associated with neuronal activity, and coupled to synaptic plasticity and to time-dependent interconnections between the cerebellum and the cerebral cortex. This calls into play a global analysis of the nonlinear space of all such maps. Combinatorial topology and field theory provide the natural mathematical framework for addressing the ensuing questions: (i) how is a specific realization of the neural functional geometry sampled out of the set of all possible configurations as the couplings vary? (ii) how can we characterize the anticipatory nature of the virtual reality generated by the brain? (iii) how can the approaches developed for cerebellar networks delving into specific representations of space-time be extended to include multiscale functional interactions in perception and cognition?

RL4 Digital Epidemiology.

Vision. Epidemiology deals with the dynamics of health and disease in human populations. It aims to identify distribution, incidence, and etiology of human diseases to improve understanding of their causes and prevent their spread. Traditional epidemiology was based on data collected by public health agencies in hospitals, doctors' offices, and in the field. In recent years, however, novel data sources have emerged, where data are collected directly from various types of digital traces, while a larger fraction of epidemiological behaviors and symptoms is stored electronically in a form accessible and amenable to analysis. Extracting meaningful information from such enormous datasets holds unparalleled potential for epidemiology.

Web-based DM is having a tremendous impact on how global health outcomes and behaviors are monitored. Online sources often provide a picture of global health that is different to that created by traditional surveillance systems. Online sources have thus become invaluable data sources for a new generation of public health surveillance systems, which operate across international borders, fill in gaps in public health infrastructures, and complement existing traditional surveillance systems. These data sources can provide valuable information about several types of disease.

Traditional surveillance methods emerged in a world unaware of epidemiological perspectives. Today, noncommunicable diseases are by far the main cause of illness and death in high-income countries. A momentous public health risk, linked to behavioral factors, is posed by cancer, along with mental dysfunctions, depression, type II diabetes, cardiovascular diseases, and pulmonary diseases. These behavioral factors are often clustered within the population; however, the processes by which the ensuing empirical patterns arise are not yet fully understood. As more individual health behaviors (e.g. diet or exposure to pollutants) and outcomes are shared online, digital epidemiology may provide an ever clearer picture of their dynamics. One outcome of this new epidemiological landscape is that individual behaviors are shifted to the center of disease dynamics and control. Individuals and groups can be studied in the rich contexts in which their lives unfold, and the relationship between disease and behavior can be studied at the level at which it actually occurs.

Studying the dynamics of disease by analyzing digital traces has great potential to deepen our understanding. To date, most data-based epidemiologic studies have focused on presumed courses of evolution, which are not yet fully established empirically. Diverse types of digital trace data may soon enhance the quality of measurements and tests. High-quality data are needed to improve the parameterization of large-scale computer simulation disease models. Introducing these models may broaden the traditional perspective, encompassing large numbers of individuals rather than population aggregates. There continue to be great improvements in how data shapes the development of computer simulation. This permits in silico experiments that are hardly feasible in real systems, providing more accurate scenarios, which are great value for policy making and crisis management.

The technical challenges of this effort are significant. Collection, storage, and analysis of massively large datasets is achieved by interfacing infrastructure, software, and algorithms. The infrastructure requirements include high bandwidth, low-latency computer networks, vast storage space, and the availability of large clusters of machines. Given the real-time, large-scale demands of data, data collection and storage software must run continuously, impervious to hardware, software, and network failures. Algorithms and infrastructure design must be efficient and scalable in order to mine, analyze and process large-scale dynamic epidemiologic data. The development of new algorithms is necessary, possibly by leveraging emerging data-processing techniques, to handle large datasets in parallel on large distributed computer systems. In addition, specific cutting-edge epidemiological DM algorithms are required to extract knowledge (e.g. filtering, classification, anomaly detection).

Cancer results from the interaction and cooperation of many intrinsic and extrinsic factors. Massive volumes of data on individual mutations, gene expression, and epigenetic factors related to cancer must be integrated into a coherent picture, in which phenotypic traits of cancer cells emerge as the result of the interactions within and between pathways. We will conduct numerical simulations of pathway response and evolution, using as input the vast body of gene expression and miRNA data. The goal here is to understand and classify all possible paths leading to cancer at the level of the entire pathway rather than focusing on single genes. Another big task will be to find correlations between nutrition and cancer, based on epidemiological data on health and diet, and to reconstruct with high accuracy the entire complex metabolic pathways needed by the cells to perform their physiological functions. Our goal here is to understand how the disease could emerge from the persistent disruption of the correct metabolic reaction in the cell; our tool will be a large-scale data-driven analysis of metabolic network evolution with a variety of diets as input.

On a different front, the etiology of sporadic neurodegenerative diseases, including AD, PD, and ALS, are believed to be based on the interaction between the genomic repertory of individuals and their environment. These hidden relationships may be studied here by taking advantage of the large dataset of patients' electronic medical records, dietary habits, and lifestyles, integrated with their biological data including genome sequences, as described in NGC.

A possible outgrowth of this data-driven approach will be to develop a computational In Silico Drug Discovery Platform in collaboration with CLSC.

RL5 Ethical issues in Data Science.

Vision. The availability of large-scale digital data is a key enabler for the Human Technopole vision. The relevant data are quite often personal data and fine behavioral data of citizens and customers. Addressing data protection is therefore crucial. The development of big data analytics for public health and predictive personalized medicine poses unprecedented ethical challenges that transcend the established vision. The desired degree of interlinking and cross-correlation of heterogeneous data sources affords the inference of extremely sensitive information about individuals. The design of population-scale and individually targeted actions based on big data analytics and behavioral science raises delicate questions on what it means to have algorithmic decision-making systems in the loop of public health, medicine, and society at large. It is mandatory to conduct foundational and applied research on how to achieve the scientific promises of BD while respecting the ethical values and legal frameworks that define our societies. This requires a long-term interdisciplinary effort that spans public health policy, medical ethics, computer science, law, and philosophy, with the goal of designing new conceptual instruments to safeguard the centrality of human beings.

We must also consider the cyber threat to the life sciences, specifically for genetic digital data within the scope of Precision Medicine and the food supply chain. Both data and interconnected software services could be tampered with, manipulated, and stolen by cyber criminals acting for economic, terroristic, or military purposes. To cope with the evolving threats, we must investigate cyber risk-management programs, cyber biotech architecture, information-sharing systems, and secure DNA data management .

C5: CENTER FOR COMPUTATIONAL LIFE SCIENCES (CLSC)

Vision: The Center for Computational Life Sciences integrates computational and bioinformatics activities to study disease-associated biological processes in order to discover innovative and personalized therapies for cancer and neurodegenerative diseases. For these activities, advanced technologies for computational simulations and big data analytics must be designed and developed. These will require efficient algorithmic and software implementations, capable of fully exploiting the unprecedented computational power of the most modern hardware architectures. Within the Center, theoretical chemists, physicists, and bioinformaticians will jointly develop novel formalisms, algorithms, and codes. Subsequently, computer scientists and software engineers will optimize these codes into more professional and production-ready software solutions, fostering their widespread international distribution and exploitation. This will also promote technology transfer activities and market opportunities. Similarly, a team of software engineers will optimize codes for High-Performance Computing (HPC) architectures, providing the end-user with flexible and efficient tools in line with state-of-the-art computing architectures. The integration of advanced big data technologies with an HPC mindset will be a major driver of Center activities.

At steady state, the Human Technopole is expected to generate huge amounts of data (from petabyte to exabyte) each year. This data will need to be efficiently stored and routinely processed for analysis. A large *HPC infrastructure* coupled to a *big data storage facility* will be a key strategic asset. It will allow the huge amounts of generated data to be efficiently processed and patient-relevant information to be extracted. To foster this key asset, a strategic alliance will be established with the Italian supercomputer center CINECA. This will further potentiate the Human Technopole's storage capability and HPC power, allowing the computational analysis of large-scale genomics data and fostering the exploration of fast-growing 1D and 3D genomics data. Computer science and computational simulations will be integrated with systems biology and network pharmacology. This will help in elucidating the genetic origin of complex traits in order to identify novel pathways, targets, and target combinations, which can be used to discover more effective and personalized medicines. All the theoretical studies will be synergistically complemented by experimental activities in medicinal chemistry and structural genomics, which will contribute to the Human Technopole's ambitious goal of revolutionizing healthcare by boosting the advent of Precision Medicine and tailored drugs. In addition, in line with HT's interdisciplinary character, this Center will closely interact with OGC, NGC, and AFNGC for a day-to-day complementarity between computations and experiments, and with DSC and CADS to develop, optimize, and parallelize new software tools for next-generation HPC infrastructures. The CLSC will comprise a central lab at the EXPO site, with collaborative programs with IIT (*RNA modeling, Genome Bioinformatics, Structural Genomics, Computational Drug Discovery*), Università Bicocca (*Computational Nanomedicine and Systems Biology*) and Politecnico di Milano (*Construction of the Storage/HPC facility*). The partnership with CINECA will be crucial for the data storage and HPC infrastructures of the F3 Facility.

Scientific Structure: From an organizational standpoint, the Center will focus on four main research lines (RL1-4). Each line will contain 3 or 4 specific activities, which ideally will correspond to an independent PI (tenure-track scientist), who will lead a group of 5-10 PhD students and postdocs. International calls will be launched promptly to recruit PIs and to activate the research activities reported hereafter. At steady state, the Center for Computational Life Sciences will comprise over 100 scientists, including PIs, researchers, postdocs, PhD students, technologists, and technicians. Notably, while all PIs will lead independent teams, it will be mandatory that they coordinate closely with each other for software development and for research activities in collaboration with OGC, NGC, and AFNGC. Furthermore, the Center will include an Outstation at the Italian supercomputer center, CINECA (HT@CINECA), which will be a partner for the entire Human Technopole's HPC and data storage infrastructures.

Research Lines and Activities:

RL1. Multiscale Modeling.

Vision. Multiscale simulations from atomistic to mesoscale levels are gaining increasing attention from the computational biology, chemistry, and physics communities. In genomic research, multiple lines of evidence suggest that the three-dimensional (3D) organization of chromatin at the kilobase-to-megabase scale plays an important functional role. DNA is an intrinsically multiscale molecule, and traditional single-scale approaches are either too approximate to grasp the subtleties involved in the complex regulation of DNA's life cycle and activity or are too detailed to account for the prohibitively large number of variables involved. Spatio-temporal interaction of DNA with proteins is a key aspect of multiscale modeling and genomics research. For example, in the near future, innovative multiscale modeling will grant a better understanding of the structure, dynamics, and assembly of the nucleosome and chromatin fibers, and their relationships with pathological conditions. In addition, multiscale modeling is emerging as a crucial tool for designing nanoparticles and medical devices on the nanoscale, and for modeling RNA, which are among these Center's other big challenges. Molecular dynamics (using classical and ab initio potentials), coarse-grained dynamics, and enhanced sampling methods will play a crucial role in computing thermodynamics and kinetics profiles associated with biological and biophysical events. Overall, RL1 will comprise four different activities, which are briefly outlined hereafter.

3D Genomics.

The 1D sequence of DNA already contains much information. However, the possibility of analyzing the 3D structures of short genomic sequences up to large chromatin fibers creates new avenues for exploiting genomic information in personalized medicine and genome-based drug discovery. The goal will be to more fully understand the genome's 3D structure and its relationship with physio-pathological conditions. This will be achieved by combining bioinformatics and multiscale simulations with experiments to better understand the genome's 3D behavior, including interactions with proteins. Research activity will need to combine high-level computational methods (quantum mechanical calculations up to coarse-grained modeling) with next-generation experiments for 3D genomics structural investigations.

RNA Modeling.

The discovery of a large number of long and short noncoding RNAs represents an unprecedented challenge for describing the structure-function relationship of RNAs. Representative examples of long noncoding (lnc) RNAs seem to be organized according to modules folded into secondary structures, with limited primary sequence homology. Computational methods for systematically identifying lncRNA secondary structure motifs are still unsatisfactory, and there are very few examples of deletion analysis to assign biochemical activities or biological functions to specific structures. This activity seeks to develop novel multiscale computational approaches to modeling lncRNAs and their interactions with proteins and DNA, with the goal of identifying classes of lncRNA domains. This research activity will also focus on optimizing these simulation tools to efficiently use HPC infrastructures and next-generation hardware architectures. Finally, these studies could create new paradigms and frameworks for designing RNA, which could be exploited for innovative RNA-based therapies. This activity will be carried out in strict collaboration with NGC.

Computational Nanomedicine.

Over the last decade, a plethora of nanodevices has been developed for the diagnosis, imaging, and therapy of a variety of diseases, including cancer and neurodegenerative diseases. However, these nanosystems have been developed following rather empirical approaches. The notion of their rational design has only recently been realized and is an innovative field of research. This activity seeks to develop novel multiscale computational modeling tools for designing and optimizing nanosystems and for predicting diseased tissue responses to nano-based treatments, with the ultimate goal of enabling intelligent drug delivery and efficient diagnosis. A further strategic topic of this research activity will be the computational design of biocompatible nanoparticles for the targeted delivery of genetic materials.

All these computational activities will be carried out in strict collaboration with experimentalists at NGC and CSMD.

Modeling of Chromatin Fibers.

The complex interplay between nucleic acids and proteins occurring at the level of chromatin underlies many fundamental physio-pathological processes, especially the control of gene expression and DNA replication. These processes must be studied at the intersection of biology, physics, and chemistry, with the use of different and complementary techniques. Theoretical methods can help produce a better understanding of chromatin's structure and function. To date, however, their practical use has been hampered by chromatin's multiscale nature and by the complexity of how its fundamental constituents interact. This research activity will focus on studying chromatin fibers, their biophysical behavior, and their role in genetics-based pathological conditions. To do this, the team will design and develop innovative multiscale computational approaches based on quantum and molecular mechanics up to coarse-grained potentials. Recent advances in chromatin modeling techniques at the mesoscopic and chromosomal scales will also be considered with a view to developing multiscale computational strategies 'from-atom-to-chromosome'.

RL2. Bioinformatics.

Vision. Novel data-intensive technologies (big data genomic technologies, in particular) are generating an extraordinary amount of information, unprecedented in the history of biology and medicine. It is not surprising that a significant part of the budget of a modern life sciences institute is today spent on information technology to handle, store, and categorize the huge amounts of data generated by experimental groups. Located at the intersection between modern life sciences and computation, bioinformatics is thus expected to greatly impact the proper description, rationalization, interpretation, prediction, and control of biological phenomena and processes. Oncogenomics, neuro genomics, and food genomics research creates specific needs for data acquisition, storage, analysis, integration, and hypothesis generation. Bioinformatics will address these needs. Within the Human Technopole, bioinformatics is expected to be central to many research programs. Several bioinformatics teams will be either integrated in Centers that perform the vast majority of experimental activities (OGC, NGC, and AFNGC), or hired as separate groups to support different research areas, including fundamental genomics, target identification and validation, disease mechanisms, and clinical research (drug-prescription databases, clinical-pathology databases, genotype-phenotype association). Innovative bioinformatics tools will be developed to be highly parallelizable and to fully exploit the most modern HPC infrastructures. In the Center for Computational Life Sciences, RL2 will comprise the four activities reported briefly hereafter.

Population Genomics.

By collecting the whole genome sequencing data (as well as whole exome, targeted sequencing, etc.) of a large number of individuals, we can correlate individual phenotypes to genetic variants on an unprecedented scale. This poses tremendous challenges at the computational level. Here, the goal is to develop new tools and paradigms for comparing large sets of genomic data and for identifying previously hidden sequence patterns associated with a disease. This team's major research activity will be to understand and compare the large amount of data arising from the genetic analysis of individuals (both nationally and internationally) affected by cancer and neurodegenerative diseases. These activities will be organized into two areas of interest: *Pharmacology for cancer genomics* will identify the somatic and germline variants that characterize each patient in a large population of individuals with different cancer types (i.e. patient stratification in oncology); *Genetic variation* will analyze the germline variants of many healthy individuals and compare these variants to those present in patients with cancer or neurodegenerative diseases. Applications of novel tools in these two research areas will help to define genetic patterns, stratify patients, and identify personalized medicines associated with well-defined genetic profiles.

Systems Biology.

Living organisms are complex dynamic systems, whose emergent behavior at the molecular, cellular, and tissue levels cannot be predicted by simply knowing the features and properties of the individual parts. This research activity seeks to integrate multiple sources of biological information (transcriptomics, epigenomics, proteomics, interactomics) and quantitative measurements of molecules and cells in human populations and in model systems. It will achieve this with mathematical models and computational analyses using data-driven strategies. This research activity will also integrate and optimize these tools for HPC infrastructures in order to generate innovative models for interpreting increasingly complex pathways, networks, and systems. Finally, systems biology activities and network pharmacology will interact to identify novel targets and pathways that could be used to discover innovative and personalized medicines.

Genome Bioinformatics.

Genomes contain the main information to build up cells, tissues, and organs. They are organized according to sequences with different functions that define structural, regulatory, and coding elements. This research activity seeks to develop computational tools to infer biological features from genomic and transcriptional sequencing data in order to unveil hidden information and functions, with an emphasis on cancer and neurodegenerative diseases. This activity's final objective will be to design and develop a new software platform of innovative computational tools for managing, analyzing, and interpreting genomic and transcriptional sequencing data. For instance, these tools could facilitate the continuous evaluation of "cancer progression models" based on cross-sectional sequence data, or a pan-cancer progression investigation to find commonalities in cancer development. Two major application areas will be considered: *Computational genomics* will interpret genomic and transcriptional sequencing data; *Computational epigenomics* will develop innovative methods for the integrative analysis of publicly available and in-house-generated epigenomics data, and their correlation with cancer and neurodegenerative diseases.

Machine Learning and Deep Learning.

This activity will focus on developing novel advanced smart tools to analyze genomic and proteomic big data. It will be based on machine learning (data mining, kernel-based learning, Bayesian networks, deep learning networks, etc.), information theory, and statistical analysis of highly dimensional data, using original methods for feature selection and prediction model building. To achieve these objectives, this activity will combine huge amounts of omics data with individual medical information (blood and bone marrow biomarkers, diet, imaging, etc.) and epigenetic factors to *i)* build prognostic models, *ii)* infer large networks of dependencies, and *iii)* identify complex patterns and correlations with patients' clinical outcomes and personalized treatment protocols. These tools will be developed and optimized for HPC infrastructures, using highly parallel paradigms (parallel machine learning, etc.) to develop codes and software suitable for big data analytics. This activity will be one of the test cases for the development of innovative and highly parallel algorithms and codes to fully exploit an HPC environment purposely conceived for Big Data. This activity will be carried out in strict collaboration with DSC for new mathematical formalisms and theories.

RL3. Software Development and HPC Optimization.

Vision. This research line will comprise the pivotal activities of the Center for Computational Life Sciences to fully optimize and exploit codes, algorithms, and software. In particular, we seek to cover the entire pipeline from new ideas, formalisms, models, and equations right through to the development of user-friendly and professional software solutions, passing through algorithms, codes, and implementations with increasing levels of sophistication. Within the Human Technopole, we will launch several informatics activities to produce innovative software for the scientific community that is ready for technology transfer and market opportunities. Computational and bioinformatics tools developed to handle the large volumes of omics data will be moved from the proof-of-concept stage to validated, standardized, highly parallel, and professional software. This will support the dissemination of best practices within an HPC/big data framework, making these tools available to the scientific community. We envision a team that translates schematic algorithms to professional implementations, designing and developing innovative software at different levels, from big data analytics to web portals and graphical user interfaces for bioinformatics and, more generally, for computational biology with a particular focus on multiscale modeling. All the new tools will be designed for optimal scalability to run on parallel and highly efficient HPC hardware architectures. These activities will also be driven by novel formalisms, algorithms, and codes generated within DSC and CADS, with RL3 serving as a hub within the structured 'equation to software' pipeline.

From Models to Software.

Based on the Human Technopole's strong commitment to technology transfer, the aim here is to establish a team of software engineers and computer scientists to develop software and new enabling informatics tools. There is a gap between the algorithms and codes designed by theoretical and computational chemists, physicists, and bioinformaticians with profound knowledge of the scientific problem, and the particularities of the newest computer architectures, which require software engineering and computer science expertise. This research activity seeks to bridge that gap. This is particularly important since a nonoptimal algorithmic implementation will hamper the proper exploitation of the huge computational power that is becoming available in the HPC community. Moreover, to realize the potential of modern computer hardware (such as for instance, General Purpose Graphic Processing Units, Intel Knights Landing, etc.), solid knowledge of the underlying architecture is required. It will thus be central to this team's mission to create a common language and to foster fruitful interactions between software engineers and the scientists from DSC, CLSC and CADS.

Code Optimization for HPC.

Algorithms and codes developed within the Human Technopole will be designed from the outset to be highly parallelizable and to fully exploit existing and future HPC infrastructures (including, at present, CPU and GPU architectures). In the meantime, existing simulation codes used within the Center must be adapted or interfaced to other software modules, which must be included in complex workflows to meet the specific needs of different research activities and to fully exploit the available HPC resources. We see HPC as a strategic and essential asset for the future of bioinformatics, big data analytics, and Precision Medicine. We have already observed that atomistic and multiscale simulations rely heavily on HPC as the research becomes more ambitious. Combining big data and bioinformatics with an HPC mindset will be one of the Human Technopole's main objectives and strengths. Code will be modernized and optimized for HPC infrastructures in close collaboration with CINECA, which has consolidated experience in HPC and code re-factoring for cutting-edge hardware architectures. Code optimization and parallelization activities will initially be carried out by using the available bioinformatics tools for Personalized Medicine programs, including GATK (for somatic variant calling), Platypus (for germline variant calling), VEP (for variant annotation), DAVID (for proteomics), and PATHIVAR (for genomic data integration).

Data Storage and HPC Facility.

The Data Storage and HPC Facility will be built in close collaboration with CINECA, which recently installed a new computing system (named PICO) to meet the ever-growing demand for services and capacity (storage, management, computing, and visualization). CINECA has extensive experience in HPC infrastructures and its latest machine (named Marconi) will reach a peak performance of about 20 Pflops by the end of 2016. A team of 10-15 research technologists and technicians will be recruited for the Data Storage and HPC Facility. They will develop specific workflows for data storage and queries of next-generation petascale to exascale databases, together with a new user-friendly web portal for these activities. These novel tools will ultimately help experimentalists and medical doctors to more easily identify biologically and medically relevant patterns. The end user will thus be provided with a professional, flexible, and modern environment to facilitate the exploitation of HPC and cloud computing in Personalized Medicine (see Facility F3 for further details).

RL4. Computational Drug Discovery and Network Pharmacology.

Vision. Huge amounts of genomics data and their correlations with clinical outcomes create new avenues for discovering innovative and personalized medicines. This research line will be fully integrated with experimental drug discovery, chemical biology, and target identification and validation activities. This research line will also include experimental medicinal chemistry and structural genomics. All the other activities covering the entire drug discovery pipeline will be conducted in collaboration with OGC and NGC. Small organic molecules, discovered using computationally-driven molecular design, will help decipher and characterize new physio-pathological pathways using chemical biology approaches. The network pharmacology activities will also be carried out in the context of the drug connectivity map, in close collaboration with OGC and NGC.

Structural Genomics.

The Human Technopole's experimental genomics activities will generate a huge amount of novel genomics information, which will in turn provide a plethora of new proteins and targets for treating cancer and neurodegenerative diseases. In this context, structural genomics will seek to characterize the 3D structures of these innovative targets for drug discovery endeavors. Structural data can accelerate the discovery of new compounds with unique and novel modes of action and binding mechanisms (allosteric modulators, disruptors of protein-protein interactions, etc.). In recent years, structural biology has been critical in directing target identification and lead discovery, with high-throughput methods for structure determination exploited as powerful approaches to screening. The ultimate goal is to identify target proteins that are relevant, innovative, and amenable to the discovery of novel medicines. The newly identified target proteins will be recombinantly expressed using different expression systems. Using X-ray crystallography, high-resolution electron microscopy, and NMR, these proteins will then be fully characterized in terms of atomic level structural information and the macromolecule's dynamic properties. Protein activity and behavior in solution will also be characterized, exploiting molecular biology and a variety of biophysical techniques ranging from fluorimetry through to dynamic light scattering (DLS) and biochemistry assays. These activities will generate new structural and biophysical information to assist and complement the multiscale simulations and computational drug discovery efforts.

Computational Drug Discovery.

To validate novel hits/leads and aid rational structure-based drug design efforts, it will be crucial to integrate computational tools with structural biophysics and X-ray crystallography. Structural data will be used in screening campaigns to assess libraries of diverse and commercially available compounds in order to identify new small organic molecules for target validation within a chemical biology framework and for lead discovery. In this context, the computationally driven drug discovery strategy, which integrates computation with medicinal chemistry and structural genomics, will contribute to the prompt and rational identification of novel anticancer small organic molecules, biological drug candidates, and new leads for neurodegenerative diseases. These new molecules will be characterized by biochemical in vitro and in vivo experiments. This will facilitate interdisciplinary drug discovery projects that rapidly move new promising therapeutics toward the clinic. To achieve this objective, we will develop and apply advanced computer-aided drug discovery protocols to design and promptly identify promising compounds that can interact with and modulate the targets generated by the OGC and NGC experimental genomics activities. Indeed, computational drug discovery will be carried out in close collaboration with OGC and NGC to identiy optimize novel candidates for cancer and neurodegenerative diseases.

Big-data-driven Polypharmacology.

This activity seeks to combine polypharmacology and big data analytics in accordance with some of the basic tenets of network theory. Indeed, when seeking to interfere with a network infrastructure whose features resemble those of pathological interactomes or transcriptomes, the best strategy is not to strike a localized albeit potent blow, but to simultaneously shut down multiple selected nodes. The main goal is therefore to develop novel drug candidates (both small organic molecules and biological drugs) that exploit, in exclusive ways, the huge amount of omics data generated and/or collected by the Human Technopole. This group will prioritize those pharmaceutical target combinations that traditional knowledge-based methods have not yet identified and that could only be identified via big data analytics. In particular, we will focus here on discovering multitarget compounds by mining omics-sized data with next-generation analytics tools, machine learning, and cluster analysis. To achieve this, we will synergize polypharmacology and systems biology to fruitfully exploit the huge amount of genomics and other omics data arising from the entire Human Technopole's experimental activities, particularly those carried out by OGC and NGC.

C6: CENTER FOR ANALYSIS, DECISIONS, AND SOCIETY (CADS)

Vision: CADS will develop original research at the intersection of computer science, mathematics, statistics, artificial intelligence, and socioeconomic sciences, endowing the Human Technopole with advanced data-handling tools and solutions. The Center is founded on the assumption that big data "don't speak" if they are not properly interrogated through original combinations of advanced analytical tools, computing power, technical expertise, and domain-specific knowledge.

The Center will process and contribute to analyzing high-throughput data gathered or generated by the Human Technopole. It will address the challenges raised by the rapid changes in data acquisition, storage, and computer processing technologies, with the primary aim of filling the current gap between data growth and processing capability. Through novel massive parallel computing algorithms and groundbreaking statistical learning heuristics, methods, and models, the Center will equip the Human Technopole to fully exploit the opportunities created by big data. Knowledge will thus be advanced through the integrated analysis of large-scale genomics data, personalized medical and lifestyle data (e.g. food and nutrition), and information continuously generated by the healthcare system. New interactive data visualization platforms will be devised for an efficient transfer of knowledge both within the Human Technopole and externally to policy makers and stakeholders.

The multidisciplinary environment of CADS, the Computational/Data Storage Facilities at HT, and collaborations with the DSC will sustain a unique effort to collect, manage, explore, and analyze large-scale and high-dimensional data on socioeconomic decisions and interactions.

CADS combines domain-specific competences/conceptual frameworks in economics and management with mathematical and statistical methods to examine variations in the state of the economy, productivity, health, wages, and education, and to develop longstanding and novel statistical indices of economic activity. Economic outcomes and effects will be measured using large-data mathematical/statistical methods (topological analysis, data mining, machine and statistical learning, predictive analytics) and econometric analysis/causal modeling of relevant decisions, activities, and relations.

Heterogeneous (and often still unstructured) public and private micro data on economic behaviors, activities, and interactions are available in massive amounts. A key overarching challenge for CADS is how to leverage the availability of this data to produce rigorous quasi-experimental research designs. Highly granular data from administrative records and private sources will be integrated and managed to study the structure of the economy and the impact of policy decisions on an unprecedented scale, speed, and resolution. This will allow us to detect targeted variations and to produce causal estimates to guide economic policy makers.

Scientific Structure: CADS will develop along 4 Research Lines (see scheme of Figure 1): RL1 HW&SW Conceptual Design; RL2 Information Processing; RL3 Modelling and Managing Socioeconomic Systems; RL4 Decisions and Policies. The CADS Center will be jointly run by Politecnico di Milano in collaboration with FEM (*Decisions and Policies*). CADS and the CADS Center will rely on the shared Facility for Data Storage and High-Perfomance Computing (with CINECA).

f

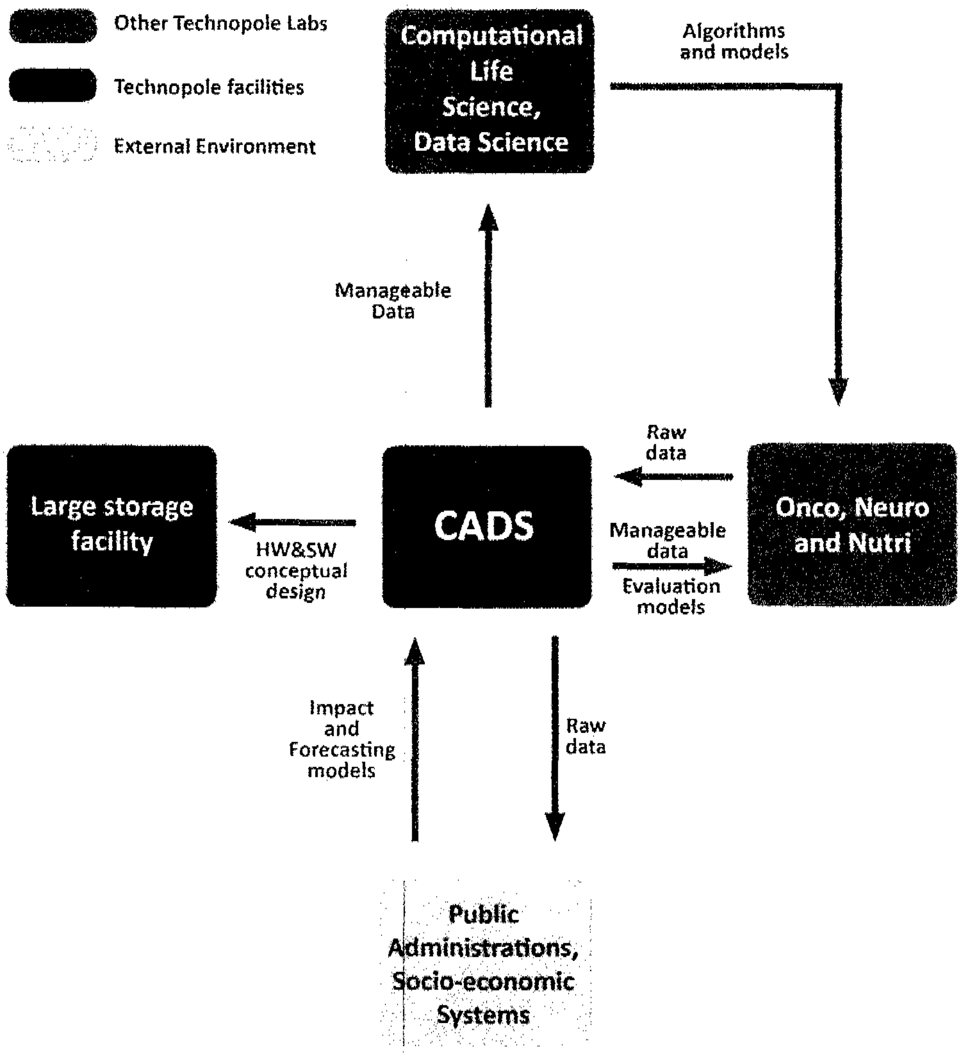


Fig.1 CADS @Human Technopole.

Research Lines and Activities:

RL1 Hardware and Software Conceptual Design.

Vision. RL1 addresses research questions and challenges in data acquisition, storage, and processing technologies. It is articulated in three main activities: i) hardware and software infrastructure research for data analytics; ii) massively parallel machine learning; and iii) embryonic technologies.

Hardware and Software Infrastructure Research for Data Analytics.

To fully exploit the opportunity created by big data, we must find ways to bridge the gap between data growth and processing capability. Hardware growth is not disruptive enough to close the gap; hence a more feasible path is innovation in software. Innovations in big data software will not only reduce the requirements of labor and expertise from analysts, but may in fact drive efficiency through more parsimonious representations of data.

Most of these solutions will require the use of high-performance computing (HPC) together with careful parallelization of algorithms, providing the ability to perform analytics in real time. HPC techniques are essential for extracting meaning from any ultra-high resolution data, such as personalized medical or stocktrading data. Parallelization can improve computation time by orders of magnitude to analyze low-frequency databases. This activity's main targets will be:

- reducing the amount of data to be processed: better compression; early detection of irrelevant t

data; and more effective sampling techniques;
- algorithmic effort/processing reduction: parallelization techniques and more efficient machine-learning algorithms to produce predictions in less time;
- improving systems effort: better utilization and sharing of available hardware resources;
- generating insightful analysis: something that produces analytics and answers at a higher level than is currently standard.

Massively Parallel Machine Learning. Big data analytics challenges the traditional paradigms of machine learning in terms of data representation, in that it typically involves a large number of heterogeneous data sources that must be used together, and in terms of computation, in that data cannot be entirely stored in main memory. Semantic analysis of text databases, forecasting of systemic risk from high-frequency economic data, and understanding patient choice and health outcomes from personalized medicine are all examples of systems which are simply too 'big' for a single computer's memory. Parallelization techniques are urgently required for a full analysis of such systems. In particular, creating practical parallelization techniques will permit the analysis of massive health and economic databases to drive a predictive analytics that can inform the choices available to policy makers. The main targets of this activity will be:

- Cloud-based parallel data analytics;
- GPU-based parallel data analytics;
- Heterogeneous parallel analytics using coarse cloud-based parallelization and finegrained in-memory GPU parallelization;
- Parallel machine learning for heterogeneous architectures.

RL2 Information Processing.

Vision. RL2 will devise new interactive data visualization platforms for the efficient transfer of knowledge within The Human Technopole and externally to policy makers and stakeholders. It is articulated in three main activities: i) data management, integration, and fusion; ii) statistical methods for complex and high dimensional data; and iii) data visualization.

Data Management, Integration, and Fusion.

Life sciences create a huge amount of data and information stored in different data warehouses, with immense potential for statistical learning. If properly exploited, managed, and integrated, this information and data may drive the evolution of epidemiology, clinical studies, and personalized medicine. The collection of heterogeneous datasets for integrated analysis is increasingly necessary to answer large-scale questions concerning, for example, citizen wellbeing and disease progression. To address such complex questions, this activity will construct methodologies to handle both the multidimensional nature of outcomes of interest and the multiscale structure of the integrated data. The overall long-term goal is to build and analyze subject-specific longitudinal data that record different features ranging from day-to-day patient records, weekly/monthly exam results, and once-in-a-lifetime genome sequencing. Since data sources are characterized by different timescales, efficient data platform methods are needed to integrate and query such widely varying data warehouses.

Statistical Methods for Complex and High-Dimensional Data.

Complex, unstructured, and high-dimensional data require novel statistical learning heuristics, algorithms, methods, and foundational theories. This activity seeks to integrate the data that is continuously generated by the healthcare system with research data, large-scale genomic and personalized lifestyle data, and medical data. The new statistical paradigms will also shed new light on big data analysis in several domains of scientific and socioeconomic interest. In particular, recent years have seen a remarkable growth in the recording of complex and high-dimensional data that exhibit a functional nature (i.e. data that can be represented by suitable curves or surfaces, with possible time dynamics). This is essentially due to the development of diagnostic devices that can provide two- and three-dimensional images and other measures of quantities of interest, captured in both time and space. Moreover, there is now an ubiquitous diffusion of apparatus that continuously collect data and communicate with each other ("the Internet of Things"). This is generating an increasing amount of complex and high-dimensional data, which can describe the evolution, in time and space, of traditional and digital human communities. Finally, high-resolution economic data is also a revolutionary force for understanding successful policies and for measuring and preventing systemic risks. The development of analytic methods and models that appropriately account for the complex nature of the data poses new and challenging problems and is fueling one of the most fascinating and fast-growing research fields of modern statistics. The overall goal of this research activity is to devise new data-driven and model-driven approaches to statistically analyze and make inferences from large volumes of highly complex data. This will include the development of advanced statistical methods for:

- functional data: smoothing, alignment and registration; exploratory data tools; regression models; parametric and nonparametric inferential tools; depth measures; methods for multidimensional functional data; methods for spatially and temporally dependent functional data; methods for functional compositional data;
- non-Euclidean data and object data;
- spatial and space-time data: models and methods for spatially and temporally dependent functional data and for object data; spatial regression models with differential regularizations; methods for data distributed over complex domains and over manifold domains; and
- multiscale analysis of multivariate longitudinal data.

Data Visualization.

There is currently a high complexity of data acquired in many different fields together with a multiplicity of sources providing data from different domains. This situation will challenge the current paradigm of modern science, which usually represents both the observation of reality and its interpretation using large and synthetic numeric matrices, tables and spreadsheets, and two-dimensional graphics. The complexity of data and of data modeling means that the Human Technopole will require innovative data visualization tools based on cutting-edge computer graphics technologies. These tools must create a fruitful synergy that embraces the rigor and methodological soundness of mathematics, statistics, and machine learning as well as the intuitions and ideas of scientists in other fields, policy makers, and stakeholders. These new visualizations will give scientists and nontechnical policy makers insights into meaningful relationships between elements of complex systems in healthcare and the economy as well as insights into the techniques used to extract these relationships.

RL3 Modeling and Managing Socioeconomic Systems.

Vision. CADS will combine topological methods, large-data statistical techniques, and econometrics to study determinants, interactions, and variations of economic data relevant to households, companies, and public actors (see Fig.2). The Center will develop novel models for the big data analysis of activities, assets, and processes relevant to private and public decisions. RL3 is articulated in three main activities.

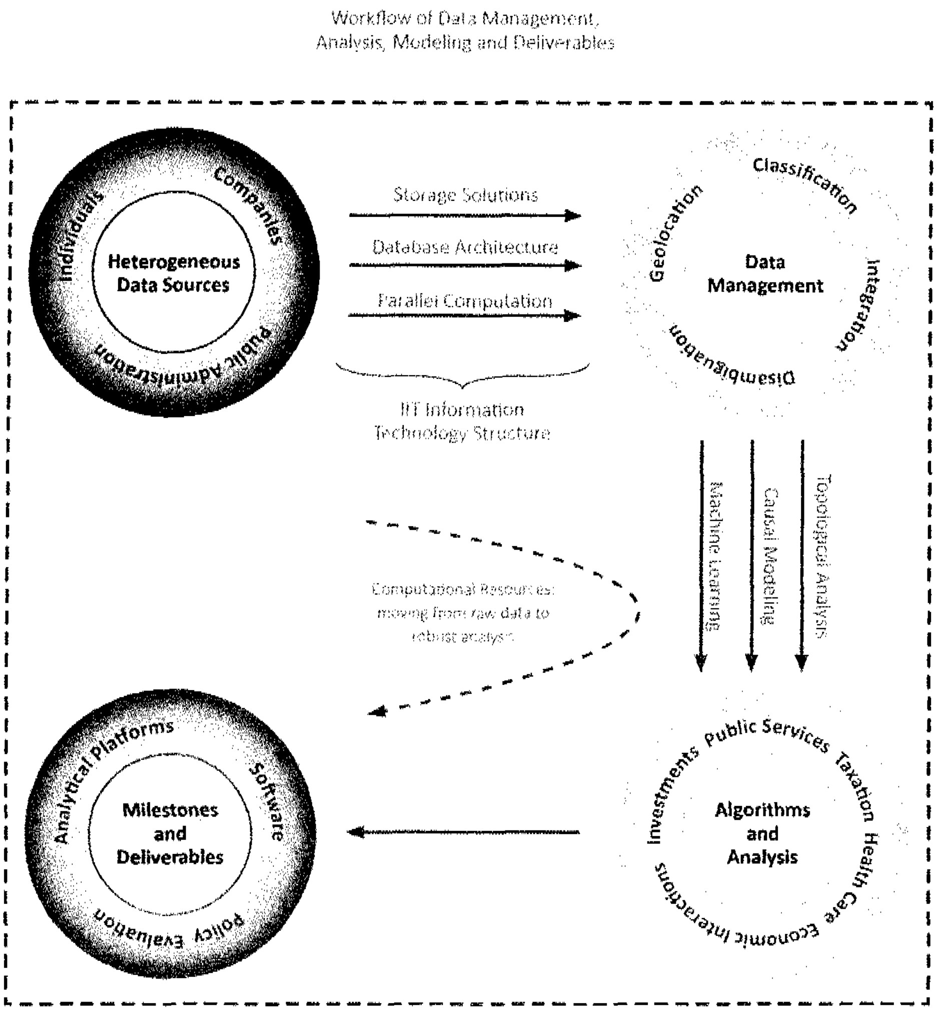


Fig.2: Analysis of socioeconomic data at CADS: Workflow.

Translational Modeling, Causal Modeling.

This activity will generate scalable methods for massive datasets in order to study the interactions between microlevel quantities and to build probabilistic forecasts of mesoeconomic and macroeconomic phenomena. Forecast frameworks at different temporal resolutions will be developed, integrating real-time/high-frequency data with lower frequency data, to address issues such as assessing the state of the economy, investigating the interactions between different groups/subsystems, and assessing the long-term sustainability of public finances in relation to aging, innovation, taxation, productivity, and growth. This activity will combine big data analytics, parallel computing, adaptive algorithms, and economic models to promote novel empirical investigations on a variety of topics. These topics include investment and pricing decisions, price fluctuations of commodities and financial products, portfolios' sensitivity to novel risk drivers, and the structure and evolution of financial and industrial networks and markets.

Performance Measurement and Management System.

A Performance Measurement System will be developed using indicators and analytical solutions based on large corpora of data from heterogeneous sources. This Performance Management System will be used to map and assess the potential benefit of the Human Technopole. More specifically, given the HT focus, research outcomes will mostly be valorized through social entrepreneurship. Interventions will be designed to ensure the full sustainability of these initiatives. In addition, the HT results must be used to engage with the public and to modify social behavior and public policy. As a consequence, we will develop communication modes, which will be accessible, effective, and understandable for the public at large. Particular attention will be devoted to developing visualization solutions and analyzing social interactions, envisaging the use of the web and social media technologies.

Innovation, Investments, and the Economy.

This activity seeks to investigate the evolution of the innovation frontier, the dynamics of industrial leadership for countries, regions, firms, and the mobility flows of individuals.

Investment decisions and location choices of companies will be monitored over time in terms of their relationship to the production of local knowledge, participation in global value chains, the resilience and sustainability of economic systems to shock propagation, and interactions with macroeconomic activity.

This activity also aims to build, maintain, and systematically update a new generation of indicators so that publicly available data from the web and private data from companies can be used to complement public indices, support a variety of microlevel studies of market behavior, and form new measures of economic activity. Researchers will analyze large-scale databases on variables such as interbank payment flows, input-output interfirm transaction, and the coevolution between prices and product attributes for hundred of thousands of units and transactions. The state of the economy will be investigated by building synthetic indicators on the variation and impact of variables such as tax collection, unemployment, consumer and investor confidence, price changes, and retail sales.

We will develop science and technology foresight observatories, while measuring the socioeconomic impact of innovation cascades in the life sciences and other relevant domains. Another RL3 goal is the innovation of assessment methodologies. We will trace the entire lifecycle of research & development by studying research processes and outputs, adoption trajectories, and their impact in real-life settings (e.g. new medical technologies in hospitals, driverless cars in complex urban environments). Researchers will investigate how scientific results and innovation outcomes are returned to society.

Finally, we will produce and maintain scalable methods, visualization solutions, and dashboards to sustain the dissemination of HT's scientific results to a variety of audiences.

RL4 Decisions and Policies.

Vision. The multidisciplinary CADS environment will sustain a unique effort to exploit microlevel real-time data in collaboration with the Data Science Center and the Data Storage and High-Performance Computing at HT. We will investigate relevant microlevel phenomena using granular real-time population-wide data in order to sustain the production of robust identifying assumptions for causal analysis.

Public Administrations, Decisions, and Society.

This activity will integrate, maintain, manage, and analyze highly granular data and detailed administrative data on individuals, public services, and corporations, with the goal of tracking outcomes from randomized experimental settings. Data are produced by national and local government agencies in the process of managing the tax system, welfare and social insurance programs, local government spending programs, etc. Timely policy design and evaluation relies on key building blocks, such as the design of randomized experimental treatment-control and quasi-experimental settings, the combination of predictive modeling and causal analysis, and the construction of systems optimized for testing and impact analysis. Population-level variations and behavioral responses will be analyzed using methods that emphasize predictive fit, deal with model uncertainty, and identify relevant low-dimensional structure in high-dimensional data, combining real-time quasi-universal population data, simulations, and experiments. We will complement causal impact measurement with large-scale data science on a variety of issues. These include manipulation of large databases, semantics and ontologies, nonlinear estimates, cross validation, methods to avoid overfitting, out of sample fit, and variable and model automated selection, Similarly, large-data algorithms in statistics and computer science will be complemented by methods to find specifications that can identify causal effects. We will focus on issues such as high-dimensional econometric modeling, the design of quasi-experimental settings, the identification of confounding variables, the introduction of instrumental variables, regression

F

discontinuity designs, and difference in differences methods. A specific focus will be on the provision of public services in domains such as healthcare and local public services. This will rely on novel composite indicators of individual attributes and preferences.

It is desirable to detect and understand, in a timely manner, variations in personal income and saving, the sources of plant and firm productivity, growth fluctuations, and so on. Achieving this requires consistency and homogenization of national, regional, local, and individual information sources, including summary statistics, accounts and subaccounts on final expenditures, incomes earned in production, and value added. This research stream will benefit from the Data Storage and High-Performance Computing Facility. It will build on the activities from RAs 1, 2, and 3 to collect, integrate, maintain, and analyze high-resolution microlevel data. These data include short-term input-output relations and composite indicators on both the income and the expenditure side of local and national accounts. We will relate these data to information on individual and household variations in wealth, income, social relations, social behavior, and individual mobility.

Health and Welfare.

This activity will generate in-depth indicators, observatories, and forecasting models on health and welfare programs and solutions. We will analyze aging-related budget items and their determinants, integrating information on healthcare, long-term care, and pensions, as well as measuring the impact of variables such as health status, career paths, retirement decisions, savings decisions, technological developments, and policy changes. We will also analyze composite heterogeneous data to inform investigations of regional and local disparities in spending and to impact the evaluation of cost-benefit analysis exercises. Hospital admission records, data on pharmaceutical prescriptions and consumption, healthcare delivery, and laboratory tests will be linked with targeted survey data and administrative data. All these data will then be analyzed. We will develop suitable methods for drawing causal inferences from observational data together with a dedicated horizon-scanning forecasting framework for pharmaceuticals.

Food and Nutrition.

This research stream aims to increase understanding of behavioral elements. It will identify the major exogenous, endogenous, hidden, and emerging behaviors that affect decisions. These behaviors should be taken into account to enhance the effectiveness and sustainability of policy, management practices, and business strategies, especially as they relate to the introduction and diffusion of novel healthier foods and food-related innovation. We will use data, a wellbeing coach, and computational models to increase information and understanding about nutrition and lifestyles. A particular focus will be on introducing novel healthier foods, food-related innovation, and promoting citizen wellbeing.

In particular, this activity aims to: i) Use big data to simulate and predict worldwide food production and consumption trends, to tailor public policies to enforce global strategies (achieve food security, improve nutrition, prevent and reduce food waste), and to stimulate local economies (promote sustainable agro-food systems); ii) Develop a wellbeing coach. This coach will act as an individual big data collector and organizer. It will also support large-scale scientific programs to monitor wellbeing, actively and concretely supporting a public awareness program on Human Technologies.

C7: CENTER FOR SMART MATERIALS AND DEVICES (CSMD)

Vision: CSMD is the Human Technopole's 'hard science' hub. Its mission is to develop material science and nanotechnologies for applications in food, nutrition, health, and medicine. The Center's mission is to deliver innovative technological platforms and devices. These include: disposable, low-cost, ultra-high sensitivity sensors for genetic analyses, food quality assessment, and food tracing; new sustainable materials for smart packaging, medicine, and other applications; and new technologies for wearable or ingestible sensors for real-time body monitoring.

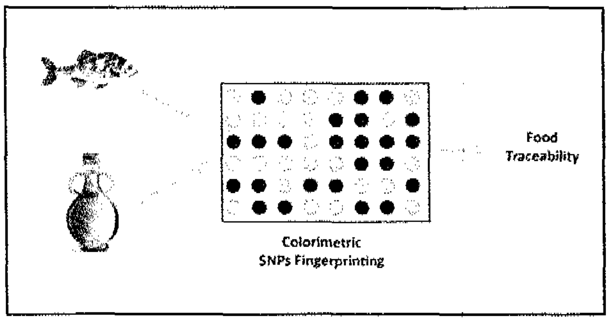
The CSMD Center will be developed by IIT and Politecnico di Milano with programs and joint laboratories with Universita' di Milano (*Human Sensing*) and Università Bicocca (*Nanotechnology for Food and Human Health, Valorization of Natural Polymers, Food Residues, and Agricultural Residues*). During the first phase, the CSMD will rely on the IIT facilities at the Center for Nano Science and Technology in Milan, providing the broad set of laboratories and skills in nanotechnology, material science, chemistry and fabrication, which the other Centers require.

Scientific Structure: The CSMD will develop 5 Research Lines: *RL1 Nanotechnology for Food and Human Health; RL2 Smart Packaging; RL3 Valorization of Natural Polymers, Food Residues, and Agricultural Residues; RL4 Water Cycle; RL5 Human Sensing.* RL1 will be dedicated to novel disposable high-sensitivity diagnostic tools for diagnostics and food traceability (OGC, NGC, AFNGC). RL2, RL3, and RL4 will be primarily dedicated to valorizing sustainable resources. In collaboration with AFNGC, these research lines will focus on agro-wastes for smart and green food packaging development, water purification systems, biomedical devices, and various other applications. RL5 will be dedicated to the real-time, wireless monitoring of biological parameters of humans, in collaboration with OGC and NGC.

Research Lines and Activities:

RL1 Nanotechnology for Food and Human Health.

Vision. RL1 will seek to develop low-cost, high-sensitivity assays for the rapid on-field testing of food and biological specimens. This research is based on proprietary IIT detection technologies, which exploit the plasmonic properties of nanostructured materials/hybrid nanocomposites and isothermal techniques. The general approach combines nanomaterial synthesis, surface (bio) chemistry, plasmon nonlinear response, molecular biology, and biotechnology. The primary focus will be the design and development of hybrid strategies, which combining the particular physico-chemical properties of nanomaterials with innovative molecular biology and biotechnology tools. These tools include DNA circuits, target recycling strategies, functional nucleic acids (DNAzymes, ribozymes, aptamers), oligonucleotide strand displacement, and cooperative hybridization. By integrating these interdisciplinary technologies, we seek to exploit the novel biophysical properties exhibited by biomolecules due to interaction with nanomaterials. These include sharp melting transitions of DNA linked to metal nanoparticles (spherical nucleic acids), the affinity enhancement by multivalency of densely functionalized nanospheres, and signal concentration on magnetic microparticles and nanoparticles. We will develop and test novel nanodiagnostics assays, including naked-eye readout detection strategies, in collaboration with the different genomics teams of the OGC, NGC, and AFNGC.

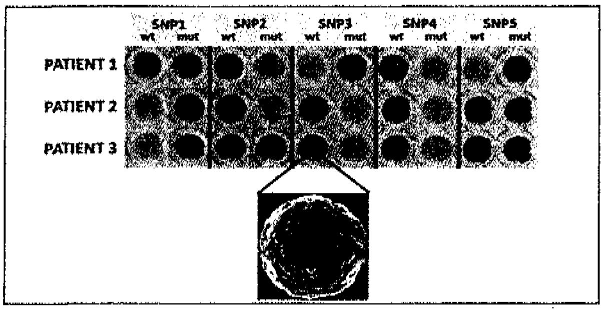


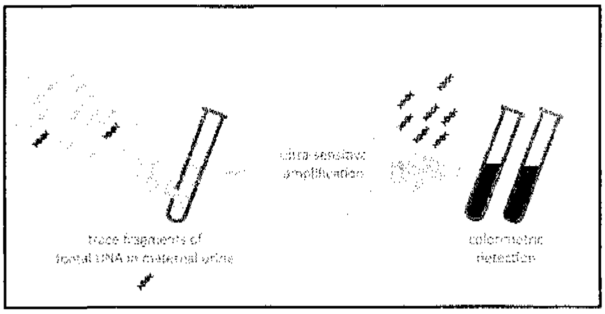
Food Safety and Traceability. Here, the goal is the early detection of food pathogens and contaminants (e.g. toxins, toxic heavy metals, pesticides) by developing rapid (10-15 min), low-cost, point-of-care (POC) tests. These will be based on isothermal (PCR-free) high-efficiency amplification schemes and colorimetric (e.g. naked-eye readout through gold nanoparticles) or fluorometric detection. They will

f

allow food processors and distributors to perform quality controls in real time, greatly benefiting both industry and public health. Moreover, we will also implement, for food traceability, tailored POC assays targeting the genetic signatures (single nucleotide polymorphisms, SNPs, fingerprinting) of particular products (plant or animal variants).

Food Characterization and Traceability. This activity seeks to create a universal (DNA-based) tool to characterize food items along the entire supply chain (from raw material to the final sold product). We will exploit methodological innovations such as the classical DNA barcoding approach and the very recent HTS techniques. This will allow us to characterize, in detail, all the components of complex food matrices and to discriminate cultivars and breeds. These DNA-based technologies will be adapted for nanodetection systems (e.g. DNA barcodes targets linked to metal nanoparticles) to provide efficient diagnostic tools for food traceability.




trace fragments of foetal DNA in maternal urine
ultra-sensitive amplification
colorimetric detection

Devices for Oncogenomics and Personalized/Preventive Medicine. We will develop low-cost and sensitive colorimetric tests (e.g. based on AuNPs) to concentrate and identify rare mutations in circulating cell-free DNA, using simplified instrumentation with naked-eye readout. The strategy will be based on a smart probe design. This will exploit the physico-chemical properties of DNA hybridization and AuNPs for increased selectivity. The enhanced detection will be due to the particular optical properties of AuNPs. These novel tests will allow early diagnosis (and early intervention) based on blood biopsies, improving clinical outcomes for cancer patients. This research line will also seek to develop alternative detection strategies for circulating oncomiRNA, relying on isothermal high-efficiency signal amplification reactions.

In the framework of personalized and preventive medicine, the primary focus will be on developing rapid and sequencing-free procedures for SNP fingerprinting (based on colorimetric (naked-eye) readout) and/or simple and effective assays, which can easily be automated and parallelized for high-throughput screening of a large number of SNPs. These simplified techniques will promote the entry of these pharmacogenomics analyses into clinical practice, greatly favouring the implementation of personalized medicine (e.g. in neuro genomics). Finally, we will seek to develop low-cost and ultra-sensitive amplification and detection strategies to enrich the foetal DNA circulating in maternal urine. These tests would offer a noninvasive alternative to standard technologies and would provide ultra-early screening of some genetic diseases.

RL2 Smart Packaging.

Vision. Here, we will develop new smart packaging technologies for food in collaboration with the Food and Nutrition Center. The goal is an active packaging technology capable of improving food preservation (through the controlled release of suitable substances) and of providing real-time in situ food analysis. We will achieve this by incorporating into the packaging materials a number of real-time sensing microdevices to detect food contaminants and spoilage in situ.

Smart Protective Wrapping and Packaging.

Here, we seek to develop sustainable packaging systems to protect food from environmental factors. One technology (called *smart wrapping)* involves surface modifications of food, using natural materials to reinforce the protection of food from environmental factors. Specifically, edible films adhere to and coat the food surface without modifying its texture or taste. The edible polymer coatings (e.g. aliphatic polyesters, natural waxes) act as protective barriers against humidity (e.g. for bread), oxidation (e.g. for fruits), and/or microorganisms. This can be accomplished using pristine polymers or by adding natural edible agents with antioxidant and/or antibacterial properties to the protective coatings. The final system (food and coating) will be fully comestible, minimizing the waste produced by the packaging. We will also develop a technology that uses multilayer packaging with controlled release of active substances from the inner layers. These multilayer packaging systems will release, in a controlled manner, natural antioxidant and/or antibacterial agents (e.g. essential oils or lysozymes) from the internal layers of the packaging to the food. The layers in contact with the food will regulate this release thanks to their special chemical or structural characteristics, while the external layers act as barriers against environmental agents.

Sensors and Indicators for Food Spoilage and Toxic Substances.

Here, we seek to develop sensors that provide visible changes in situ when gases are released during food spoilage or when pollutants are present. We will develop packaging materials with integrated sensors and indicators to detect and signal food spoilage and/or the presence of toxic substances (e.g. pesticides). Specifically, we will incorporate active molecules (e.g. anthocyanins and photochromic componds) into the natural or synthetic polymers of the packaging system. These molecules will change state (e.g. color) in the presence of pollutants or gases released during food spoilage (pesticides, toxins, metal ions, biogenic amines, acid gases, CO_2). We will design the various composite systems to be fully edible, activated on demand, highly selective, separately or in combination. It is crucial to select appropriate indicator molecules with specific affinity for targets. We will use experimental and computational models to support this selection.

Integrated Organic Electronics for Sensor Control cnd Communication.

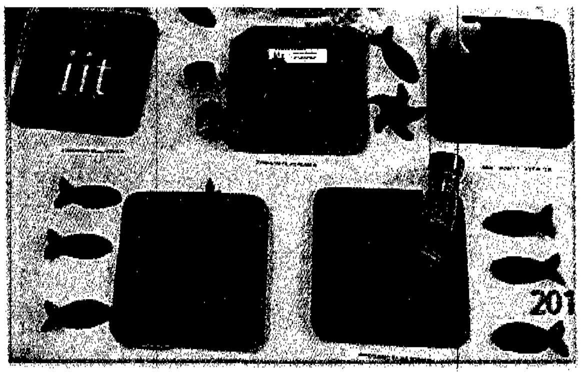
Here, we seek to develop new self-powered labels, which are written directly onto the food packaging and which embed self-powered electronic components. Organic electronic devices can be realized on mechanically conformable and compatible substrates that are easily integrated into food packaging. It is possible to print large volumes of components at very low cost and fully integrated into the packaging production line. A 'component' here is defined as a functional unit with sensing, processing, and communication capabilities. Depending on the requirements, these devices can be very simple and totally passive (like antennae integrating a sensor) or more complex structures that comprise a power source, a logic circuit, a sensor, and a transmitting device.

RL3 Valorization of Natural Polymers, Food Residues, and Agricultural Residues.

Vision. The accumulation of unmanaged agro-waste increases environmental concerns, mainly due to the high disposal costs. Transforming food waste and agro-waste into sustainable, energy-efficient materials is a viable way of reducing pollution and conserving natural resources for future generations. The RL3 objective is to valorize food waste, agro-waste, pure natural fibers and polymers to develop easily scalable methods for producing fully biocompatible and sustainable biomaterials and plastics.

Composite Bioplastics from Natural Fibers, Food Waste, and Agro-waste.

Here, we will explore innovative processes to develop composite bioplastics from natural fibers, food waste, and agro-waste. Natural fibers (e.g. cellulose) will be combined and modified with very low quantities of biocompatible macromolecules, synthetic polymers, natural polymers, or fillers. The resulting moldable composites can then be processed using the injection or extrusion techniques typically used to process plastic materials. We will form composite biopolymers by using the synthetic or natural polymers in small quantities as binding and functionalizing agents between the food and agro-waste microparticles. We will adapt and develop proprietary IIT technologies to facilitate this. Furthermore, by



precisely controlling the polymerization/modification process at the molecular level, the raw food and agro-waste materials can be transformed into high-performance plastics. Contemporary polymer chemistry provides several controlled polymerization techniques (RAFT, ATRP, ROP, ROMP, and click reactions), which can be used to exploit the various potential functionalities of a bio-derived raw material. We envisage two approaches, depending on the raw material's chemical structure. These are the modification of natural polymers (typically cellulose and lignin) and the polymerization of renewable monomers (terpenes, fatty acids from plant oils, lignin-derived monomers). Here, we will develop scalable procedures for functionalizing lignin or cellulose using RAFT agents and growing elastomeric polymers on them using vegetable oils. We could thus produce new materials with tunable thermomechanical properties. The developed bioplastics could find application as passive or active packaging systems. For active packaging systems, we will valorize the natural antibacterial and antioxidant properties of food waste and agro-waste (e.g. cacao husks, carob powder, redbeet root, orange peels, cinnamon, and lemon peels). Alternatively, we could modify the the plastics at the molecular level to achieve O_2 barrier properties. Specifically, we would modify the structural properties of the polymers with appropriate chemical functionalization (acetylation, grafting) or by incorporating food-compatible fillers (clays, halloysites, nanocellulose). As above, a particular focus will be to develop scalable methods to process the formed bioplastics for easy industrial application. We will optimize our results by fully exploring the type of processing used (melt extrusion, injection molding) and by appropriately modifying the properties of the bioplastics.

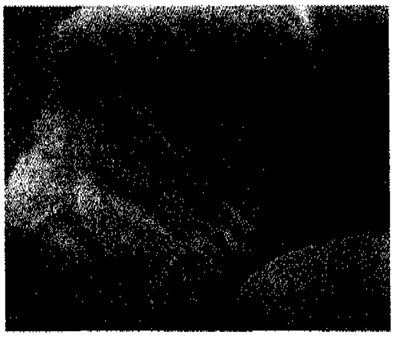
Waste Valorization by Biotransformation.

Here, we will collaborate with the University of Milano-Bicocca on the microbial upgrading of waste into desired chemicals. Knowledge of the microbial world's potential to metabolize molecules must be integrated with systems and synthetic strategies for constructing engineered microbial strains or consortia. Indeed, this integrated approach has already led to increasing numbers of successful waste valorization examples. We will use organic molecules from agro-waste as feedstock for microbial bioconversions. Selected microorganisms will be then tailored for sustainable growth and the production of goods and products using food waste. These products could find application as nutraceuticals, chemical platforms, biocatalysts, biofuels, and biomaterials.

We will explore microbial biodiversity by considering particular physiological and metabolic traits (tolerance to extreme temperatures, pH values, uncharacterized enzymatic activities, variety of substrate degradation ability), as well as the assembly and annotation of the novel genome. This would include the possibility of deciphering the nature of microbial hybrids, a trait highly diffused in industrial strains. Illumina sequencing and Pacific Biosciences technology, which produced long reads, will be pivotal here.

The engineering approach(es) will seek not only to streamline the carbon flux to the different products of interest, but also to improve the stability and robustness of the microbial strains. We will therefore apply synthetic biology principles to *i)* potentiate the power of metabolic engineering to extend the cellular rewiring from the specific pathway to indirectly connected networks in order to attain the desired final product, and *ii)* evoke a cellular response to meet industrial requirements and address the uneven nature of waste biomasses with microbial physiology and catalytic activities.

As a complementary activity, when direct approaches are not expected to be trivial or when identified strains are difficult to manipulate, we will use reverse approaches combined with screening protocols. Omics analyses can then be used to decipher the results and to further implement strain development.



Some of the chemicals extracted by these processes could be used as monomers to produce new polymeric materials with the above-mentioned polymerization techniques. Among the chemical platforms, organic acids are a key group of building blocks that can be produced by microbial processes. Most are natural products of microorganisms, or at least natural intermediates in major metabolic pathways. Indeed, because of

F

their functional groups, organic acids are extremely useful starting materials for the chemical industry, with prominent examples including succinic acid and lactic acid.

We will develop the production process, starting from lab-scale shake-flask cultures and scaling up to bioreactors, where parameters are settled and optimized for industrial production. From the outset, the development of strains and processes will be led by downstream processing and final use requirements.

Biomedical Materials from Food Waste and Agro-waste.

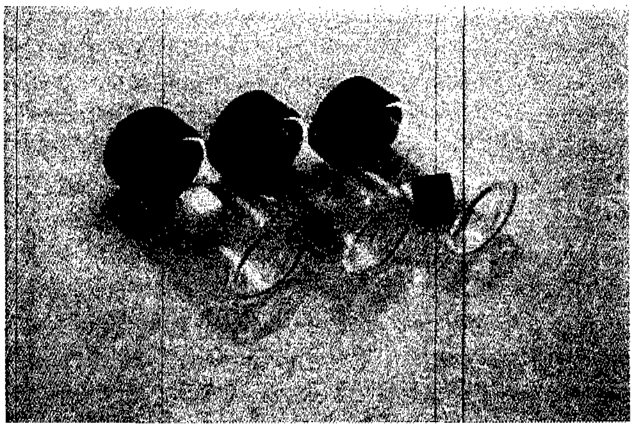
Biopolymers formed from food waste and agro-waste can be used to make pharmaceutical devices and products, cosmetics, and wound-healing materials, which all exploit the abundant vitamins, antioxidants, anti-inflammatories, proteins, and/or antibacterial molecules in these systems. As such, one interesting strategy for developing waste-based medical and cosmetic devices is to selectively extract molecules from agro-wastes and to process these to produce, among other things, nanofiber mats, films, and hydrogels.

RL4 Water Cycle.

Vision. Here, we seek to develop novel nanotechnologies for water analysis and purification, in collaboration with AFNGC and CREA. We will develop activities based on proprietary IIT technologies.

Indication and/or Removal of Pollutants inWater.

Porous materials (foams, filters, membranes) will be developed with engineered surface properties and appropriate morphological characteristics in order to remove oil from water (oil-water mixtures, emulsions). The focus will be on designing the porous support and the appropriate surface treatment with polymer films, particles, or combinations. We will also seek to develop porous composites of synthetic or natural polymers combined with natural fillers (e.g. agro-waste particles or their derivatives, including DNA). These materials should be able to interact and entrap metal ions or organic pollutants from water (e.g. coffee for lead and mercury, orange peel for nickel and polar dyes). Here, the main goal will be to develop methods for preparing the composite porous systems and identifying the appropriate filler for each pollutant, based on experimental and theoretical modeling data.



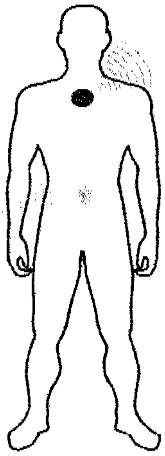
In more specialized applications, we will functionalize the developed solid materials to change in color, toughness, size, or conductivity in order to indicate the presence of organic pollutants (pesticides, dyes, amines, medicines etc.) or inorganic pollutants (e.g. heavy metal ions). The developed materials should have a fast response, low detection limit, and high selectivity towards specific pollutants.

RL5 Human Sensing for Real-time Physiological Monitoring of The Human Body.

Vision. Here, we will collaborate with the Neuro Omics Center to design and develop ingestible/edible devices for human sensing inside the body and for advanced healthcare. This new generation of ingestible electronic devices will be used to monitor patient compliance, for diagnostics, and for precision therapies. The devices will integrate IN/OUT communication systems, a power source, and sensing stages for biomarking on a flexible, biocompatible, and biodegradable platform.

Fabrication and Testing of Ingestible/Edible Devices.

We will systematically screen candidate materials for use in devices such as transistors and sensors. The substrate comprises 99% of the device mass and will be made of naturally occurring materials (silk, shellac, cellulose) that are fully bioresorbable and/or biodegradable. Similar materials could be used as dielectrics in field-effect transistors, where other choices could be genetically engineered peptides, developed in collaboration with OGC and NGC. Organic conjugated materials can be used as semiconductors. Inert metals in less than toxic levels can be used for electrodes or conducting polymers like PeDOT:PSS. Encapsulation can be achieved using standard materials tested for food industry, such as bovine gelatin or caramelized glucose. The impact of the ingestible devices on real-time diagnosis, control, and therapy can be enhanced using an onboard communication system to provide access to information and eventually receive instructions from outside. This system will be integrated into the edible device platform. Simple systems might be designed as passive RFID reading antennae. More advanced active systems will be able to transmit elaborated information and receive information from outside. This will require active circuitry, powering, and wireless IN/OUT communication technology. We will therefore develop and integrate energy sources into the ingestible/edible device to power the reading, transducing, and communication functions. We will use edible and very low-capacity galvanic cells, which can exploit the stomach acid environment, to design and develop novel suitably encapsulated and miniaturized batteries, which are nontoxic and ingestible. We will also investigate other energy-harvesting systems based on nanoactuators. Sensors will be integrated into the ingestible/edible technology platform. Sensors will be sensitive to local oxygen, pH, temperature, and/or specific substances or microbial populations. This requires soft lithographic approaches to the device's design and biochemical functionalization, as well as a proper interconnection to the device's signal transmission module. A complete device will include IN/OUT communication, an onboard power system, and a sensing stage. Such a device could offer an ideal system for local drug delivery on demand.



F

F1: CENTRAL -GENOMICS FACILITY

Scientific Structure: The Central -Genomics Facility will gather the *"Common Technology-Development Platforms"* common to OGN, NGC, and AFNGC and the high-throughput *Next-Generation Sequencing infrastructure* for large population screenings.

Common Technology-Development Platforms.

Vision. As the common technological core shared between the OGC, NGC, and AFNG Centers, we will develop, set up, and implement the essential omics technologies required by these centers. The F1 Facility will comprise specialized technological platforms run by dedicated technologists (highly specialized technicians), and coordinated by teams of Principal Investigators from the three centers. This structure will allow each technological platform to be developed and adapted to meet the scientific needs of the OGC, NGC, and AFNG. Fully validated techniques and pipelines of broad applicability will be transferred from the technological platforms to the HT Facilities and national partner institutions for their immediate use. This technological core will also serve as a training center for highly specialized technicians of partner institutions in Italy.

Genomics Platform.

The Genomics platform will set up the initial sequencing technologies and develop new genomics technologies in collaboration with the academic groups. Its portfolio will include established sequencing technologies such as genome sequencing (whole genome and whole exome sequencing), transcriptome sequencing (RNAseq, small-RNA sequencing), chromatin immunoprecipitation-sequencing (ChIP-seq), gene-panel sequencing, as well as more specialized sequencing approaches (RepliSeq; DNA-damage mapping; CAGE libraries preparation and sequencing). The platform will use the main technologies available, including Illumina, Ion Torrent, Nanopore, and Nanostring. Importantly, we will continuously seek out novel emerging genomic technologies to identify and test the most promising and economically advantageous approaches. One focus will be on developing new sequencing technologies, including single-molecule and single-cell sequencing. We will also periodically assess the potential clinical implementation of these approaches. Some of the NGS technologies initially set up here will be transferred/adapted to the HT's Large-Scale NGS Facility (F1). We will also set up standard operating procedures to establish collaborations with industry and to test various platforms.

Functional Genomics Platform. The Functional Genomics platform will provide the set-up, instrumentation, tools, technological know-how and development required for low-to-medium-throughput genomics screenings. The platform will be organized in three main activity areas (Genetic Screens, Chemical Screens, and Antibody Screens) in close collaboration with the scientific teams. External collaborators will complement internal know-how, providing additional tools to optimize performance and secure continuous technological updates and state-of-the art experimental workflows. *a) Genetic Screens.* Here, we will set up robust high-quality biological assays for medium-to-high-complexity genetic screens using pooled shRNA or sgRNA libraries (in vivo patient-derived xenografts - PDX, organoids, primary cells, established cell lines). Critically, we will obtain primary patient material via collaborations with strategic partner clinical institutions. We will seek to generate PDX models of human cancer to address the most relevant clinical needs (e.g. metastatic tissues from patients following one or more lines of therapy). Functional studies will be complemented by full molecular characterization of the patient tumor/matched PDX, including genome (whole exome), epigenome (DNA methylome and histone marks, ncRNAs), and transcriptome. *b) Chemical Screens and Antibody Screens.* Here, we will seek to: i) identify appropriate chemical or antibody screening strategies for validated targets generated by genetic screenings or academic research projects; and ii) execute small-molecule-based and recombinant-antibody-based screenings and identify confirmed hits for further development. We will collaborate closely with academic scientists to decide on the appropriate targets or screening strategies. These scientists will provide critical know-how on targets and screening assays. For small molecule screens, we plan to perform virtual, biochemical, and cellular screens. We will use commercially available libraries of different sources (30,000-50,000 compounds

F

or more; small libraries of clinical-stage compounds; natural compound libraries). Our focus will be on screen strategies against targets previously considered "undruggable" (e.g. transcription factors). Antibody screenings will be performed in vitro (against model peptides such as tumor-specific proteins), on cells, or in vivo (e.g. on intact human tumors maintained as xenografts) by phage-display technology using high complexity (>10^9 molecular diversity) human recombinant antibody libraries. To identify therapeutic targets, we will apply genome editing and conventional mass spectrometry strategies to the best candidate antibodies.

Functional Proteomics Platform.

Here, novel biochemical methods for sample preparation will be integrated with state-of-the-art high-resolution mass spectrometry technology, based on the most recent hybrid quadrupole Oribtrap Q Exactive, which allows high-accuracy MS1 scans and ultra-fast MS/MS sequencing rates. MS instruments will be equipped with HPLC systems operating at high pressures (>10 000 psi) and with longer columns (50 cm) packed with small particles (<2 μm). This LC-MS setting will circumvent pre-fractionation steps, minimizing sample loss and enabling the analysis of very low-abundance protein samples (e.g. primary cells or laser-micro-dissected patient samples). The shotgun analytical workflow will be complemented by targeted MS proteomics. Here, we will profile a pool of candidate proteins (selected during the initial discovery-phase MS screening) in a multiplexed fashion in hypothesis-driven analyses. This application will be driven by parallel reaction monitoring (PRM) by a hybrid quadruple-Orbitrap MS analyzer. The platform will incorporate informatics and bioinformatics tools for statistical and functional data analysis, including integration with data from other omics approaches (transcriptomics, genomics, ChIP-seq). It will also include a panel of biochemistry and molecular biology tools for follow-up studies. Overall, we will seek to: *a)* integrate proteomics into the screening/drug discovery process to boost the elucidation of the mechanisms of action of drugs/chemical-probes in an unbiased and quantitative manner; *b)* execute continuous proteome/modificome profiling upon perturbation of cellular processes (through drugs or genetic tools); *c)* perform qualitative and quantitative assessments of complete proteomes in biologically relevant samples, from cultured cell lines up to primary cells, fresh, frozen, and FFPE tissues; *d*) execute proteomic studies to identify novel biomarkers. This will be based on the proteome-based epigenome mapping of patient samples and on the global analysis of post-translational modifications (phospho-proteome, acetyl- and methyl- proteome); and *e*) conduct systematic MS analysis of global interactomes, including protein-protein, protein-nucleic acid (DNA,RNA), protein-modification, and protein-probe interactions. The goal will be to assess their plasticity in basal and disease states, upon pharmacological treatment, or upon any other perturbation of the model system.

Mass-Cytometry Platform.

State-of-the art mass cytometry is a recently developed technology that allows analysis of individual cells in heterogeneous populations. It can simultaneously detect up to 100 parameters on single cells, using antibodies tagged with stable isotopes of transition elements. We will focus on developing: i) new computational techniques to analyze the massive and high-dimensional mass cytometry datasets; and ii) new mass cytometry applications to analyze formalin-fixed paraffin-embedded tissues (imaging mass cytometry).

Metabolomics Platform.

The Metabolomics platform seeks to detect and quantify small molecule metabolites and lipids (using targeted and nontargeted approaches) on a variety of biofluids, cell/tissue extracts, and microbes. Initially, the platform will be equipped with high-resolution mass spectrometry coupled to chromatographic systems (LC-MS and GC-MS).

Large Next-Generation Sequencing Infrastructure.

Vision. The facility will provide high-throughput next-generation sequencing for the large population screenings planned by OGC, NGC and AFNG, and for other nationwide screening initiatives that the HT will eventually launch. This large national sequencing infrastructure will implement robust pipelines of highly standardized processes for large numbers of samples. The basic experimental workflow will comprise library preparation and sequencing with ad hoc variations to address the specific aims of the

Centers. Library preparation, fragmentation, and size selection will be automated. We will provide raw-sequencing data and primary analyses (fastq files and quality-controlled information) to HT Centers for further analysis. Activities will initially include RNA sequencing, targeted exome sequencing, and whole genome sequencing. The facility will have dedicated personnel and infrastructure comprising several laboratories equipped with HiSeq4000 sequencers (or equivalent) and with a HiSeq x Five (or equivalent). The HiSeq 4000/5000 instruments are state-of-the-art DNA sequencers, capable of running two independent 8-lane flowcells at a time (reads can be 50 to 100 nucleotides long, with up to almost 5 billion reads, 750 GB per run).

The Facility will later operate other technologies, including those optimized by the HT Genomic Development Platform. One example is the single-molecule real-time sequencing platforms, similar to the PacBio RS II instruments (or equivalent) with high-performance optics, automated liquid handling, proprietary SMRT cells, and reagents. This technology allows the identification of haplotypes and splicing variants by single-molecule sequencing and the precise mapping of repetitive elements in long reads.

F2 STRUCTURAL IMAGING FACILITY

Vision: The past two years have seen a revolution in structural biology. Single-particle cryo-electron microscopy (cryo-EM) was once limited to determining structures at resolutions so low that chemical features could not be distinguished. But with the advent of direct-electron detectors, single-particle cryo-EM has begun to achieve atomic resolutions that were previously available through crystallography only (Fig. 1). At these resolutions, chemistry can be related to both structure and sequence.

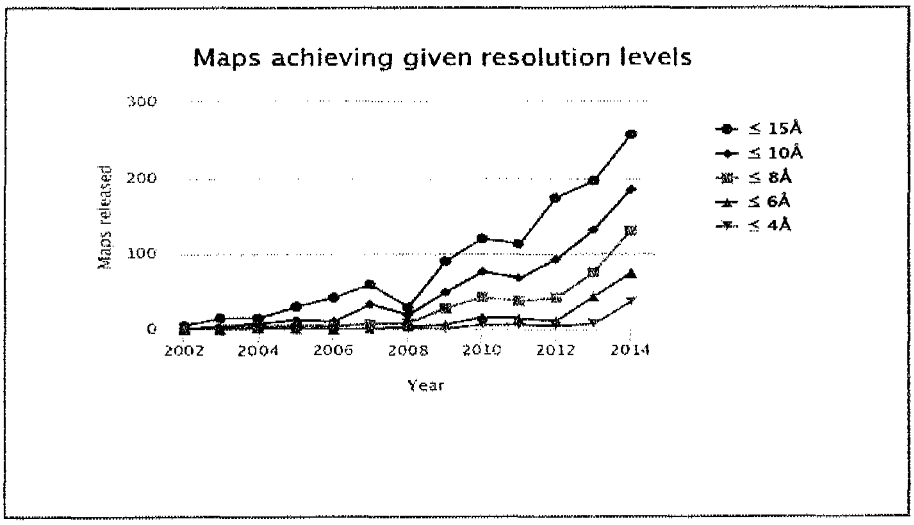


Fig. 1 Improvement of the resolution level achieved in the last years

Importantly, cryo-EM does not require crystal formation and needs only minuscule amounts of homogeneous sample. In principle, any biological problem can be tackled, with sample preparation (i.e. biochemistry) being the primary (if not only) limiting factor.

This technique is thus mandatory for a comprehensive approach to Precision Medicine. The genome sequencing of each individual patient will provide us with information on the mutations underlying genetic diseases or cancer. Meanwhile, structural proteomics can provide additional information, namely: (i) Why does the mutation cause the disease? (ii) How can new drugs be designed for currently untreatable diseases? (iii) What existing drugs might be useful for known diseases? (iv) How can existing drugs be optimized to increase their power?

Scientific Structure: To place the Human Technopole at the forefront of innovation in biomedical research, a state-of-the-art cryo-EM infrastructure is needed. This infrastructure should be available to a large group of structural biologists, who are already proficient in the biochemistry relevant to sample preparation. At steady state, this Facility will comprise labs for Electron Diffraction Tomography and Electron Cryo-Tomography.

Electron Diffraction Tomography (EDT).

Vision. A comprehensive approach to structural biology must involve the study of protein crystals, when obtained. If the protein crystal is large enough to diffract under a synchrotron X-ray beam and to withstand radiation damage under cryo conditions, X-ray diffraction is the mandatory choice. If the crystal is too small to meet these criteria, the only X-ray-based alternative is femtosecond diffraction at free electron lasers (XFEL). However, electrons can offer another alternative. Electrons interact more strongly with matter than X-rays and they deposit 2-3 orders of magnitude less energy into a crystal per useful scattering event. This means that electron diffraction can be used to investigate protein crystals smaller than 1 micron. In the past, these studies were carried out by accumulation of single diffraction patterns, allowing the structural determination of 2D crystals of membrane proteins only. However, an automatic data collection method, based upon crystal rotation (electron diffraction

F

tomography, EDT) and originally developed for inorganic crystallography, was recently applied to protein crystals (MicroED). Following this approach, 3D sets of electron diffraction data were collected successfully in cryo-EM conditions on lysozyme and catalase samples, and the crystal structure of the toxic core of α-synuclein protein was determined to 1.4 Å resolution (Rodriguez et al. *Nature* 2015 525:486). The crystals analyzed are invisible under an optical microscope and too small to be investigated on standard macromolecular synchrotron beam lines. These samples would therefore have been considered a failed experiment had EDT not raised them to the level of feasible science. Although still in its infancy, we expect this technique to be receive a strong boost once the recently developed direct electron detectors specific for diffraction (Timepix detector) are widely used to collect protein data. These detectors will allow extremely fast data collection times under low-dose conditions. Today, this approach is at the same stage as single particle cryo-EM was only three years ago. Having both capabilities available will be a unique opportunity for the Human Technopole.

Electron Cryo-Tomography (cryo-ET).

Vision. Electron cryo-tomography now offers the possibility of studying cellular structures in both physiological and pathological conditions at an unprecedented level of detail.

Tomography provides lower-resolution information compared to other techniques (such as cryo-EM or X-ray crystallography), but it allows a full 3D reconstruction of a single defined but large (multimolecular) object by recording a series of views. For example, the method can provide mechanistic details of nanoscale cellular processes such as cargo transport, virus uncoating, membrane modulation, and membrane trafficking. It can reveal a new landscape of previously unknown details of cell structure, allowing the visualization of macromolecular complexes in their native environment and host–pathogen interactions at the supramolecular level. Until recently, a major limitation of cryo-ET was that only the thinnest (<<1 micron) regions of cells could be studied intact. However, "specimen thinning" by cryo-sectioning or ion beam milling can now overcome this limitation.

For 3-D reconstructions of cellular structures or macromolecular assemblies within cellular structures by cryo-ET, the major limitations come from the conflicting requirements of a low-electron dose (to minimize radiation damage) and a higher dose (for a sufficient signal-to-noise ratio). On a practical level, a full tilt series of each object must be collected with repeated electron exposures. Each tilt exposure must have sufficient signal for alignment. Overall, this results in a high cumulative dose. These requirements and limitations currently restrict cryo-ET to nanometer resolution. However, the technique is playing an increasingly important role in defining cellular complexes and assemblies and in extending the reach of structural biology from the cellular level to the molecular level. Interestingly, in order to determine the molecular structures of subcellular assemblies, a useful and increasingly popular approach is to image the same structures with both cryo-ET and fluorescence microscopy. Indeed, the rapidly expanding power of fluorescence microscopy is ideally suited to identifying areas and events of interest within cells. These can then be examined in molecular detail by electron imaging via cryo-ET.

Convergence and Synergy with the HT Centers and Facilities. F2 will act as a Structural Biology hub: molecular targets will be identified by the sequencing Facility within the context of the activities of the Oncologic, Neurosciences and Nutrition Genomics Centers. These targets will then be sent to the Electron Microscopy Facility for structural characterization. This workflow will be of paramount importance in exploiting protein targets for drug design and in characterizing the molecular mechanisms of oncological and neurodegenerative disorders. The Structural Imaging Facility will collaborate closely with the Big Data and Biocomputing Center. This is crucial as the Facility will easily produce more than 1 Terabyte of data per day, creating a data analysis challenge that can only be addressed by the Human Technopole's integrated multicenter structure. The Structural Imaging Facility will also stimulate the development of original methods for the 3D reconstruction of molecules and cellular substructures.

Organizational Plan and Instrument Time Allocation. The Facility will be directed by an internationally established scientist, hired by international call, with a staff of 2-3 junior PIs (tenure track and/or technologists). This team's in-house research activities will occupy up to

F

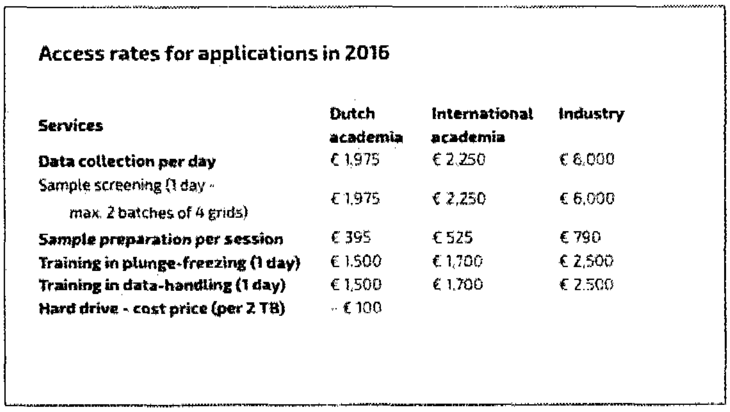
25% of the Facility's machine time. Rapid-turnaround peer-reviewed applications will be used to allocate 50% of machine time to the Human Technopole Centers and to the wider Italian and European structural biology community, with prices and policies to be established.

The remaining 25% of machine time for the cryo-EM and cryo-ET microscopes will be made available for proprietary research to companies, as already implemented for synchrotron sources.

We will actively pursue participation in European and global initiatives (Instruct, Inext) in order to position the Facility within a broader international context.

Strategic Positioning. The Facility will be strategically positioned within the Human Technopole's scientific and technological structure in order to achieve convergence and synergy. Its geographic location is well-suited to supporting Italian, European, and global scientific and industrial enterprises. Its loction is readily accesible for sample shipments, staff secondments, and data collection trips. This is in contrast to the somewhat secluded locations that are typically forced upon synchrotron sources by building and size requirements.

Equipment. In Europe, there are national facilities in Oxford and Leiden, which can be accessed upon upon application to iNEXT-eu.org (Instruct like). However, this access is limited as only a few slots are likely to be available. Nevertheless, iNEXT-Instruct is a precious opportunity to explore for training and for feasibility studies. The below table details the access rates for NeCEN for academia project submission, with peer review being a prerequisite for access.



Access rates for applications in 2016

Services	**Dutch academia**	**International academia**	**Industry**
Data collection per day	€ 1,975	€ 2,250	€ 6,000
Sample screening (1 day - max. 2 batches of 4 grids)	€ 1,975	€ 2,250	€ 6,000
Sample preparation per session	€ 395	€ 525	€ 790
Training in plunge-freezing (1 day)	€ 1,500	€ 1,700	€ 2,500
Training in data-handling (1 day)	€ 1,500	€ 1,700	€ 2.500
Hard drive - cost price (per 2 TB)	~ € 100		

Other laboratories around the world include: Max-Planck Institut, (Martinsried, DE), Medical Research Council Laboratory of Molecular Biology (Cambridge, UK), Leiden University & NeCEN (The Netherlands), NY Structural Biology Center (New York), UCSF (San Francisco, CA), Baylor College of Medicine (Houston TX), University of Toronto, Tsinghua University - National Center for Protein Sciences (Bejing), and the National Center for Protein Centers (Shangai).

Market. Big pharmaceutical companies (e.g. Merck) are in the process of setting up their own facilities. Medium and small pharmaceutical and/or biotech companies will instead continue to seek access through other companies (http://www.nanoimagingservices.com CA, USA), noncommercial facilities (see NeCEN above), or, for example, a commercial facility being set up in Strasbourg (NovAliX). A joint FEI-LMB facility is envisaged in Cambridge (UK). The main issue for pharmaceutical companies is proprietary data and data access, which often creates conflicts when companies seek to use instrumentation belonging to universities or other nonprofit entities. With the revolution in resolution achieved by cryo-EM, the interest from pharmaceutical and biotech companies is huge and rapidly growing. This is because the structures of proteins and their complexes, which were not previously amenable to crystallization, are now accessible or almost within reach.

Laboratories and Infrastructure. The Facility could be endowed with one top-range electron microscope for single-particle cryo-EM with a direct electron detector and a screening microscope. Nevertheless, the facility should have the following scalable infrastructure, so that later requests for cryo-EM or cryo-ET capacity can be met:

- High-resolution cryo-EM 300kV TEM with Cs corr, phase plate, and K2-direct electron detector.
- Screening TEM (200kV with autoloader) with different detectors (direct detector and detector for fast screening).
- Image processing, accessories, Multi CPU cluster.

The facility requires an electrically insulated and vibration-insulated site, with a humidity-controlled room.

At a later stage, HT could consider setting up the Tomography facility with:

- Tomography Cryo-EM Krios 300kV TEM with phase plate and K2-GIF.
- Cryo-dual beam SEM for Focused Ion Beam (FIB).

F3 FACILITY FOR DATA STORAGE AND HIGH-PERFORMANCE COMPUTING

Vision: Population-wide genomics data is changing medicine's central paradigm, defining pathology and associated treatments at the level of individuals. Personalized Medicine promises a quantum jump in our ability to fight illness, opening up unexplored frontiers for treating complex diseases. The field is evolving rapidly, fueled by flagship projects generating hundreds of Petabytes of data each year. These heterogeneous data (omics, clinical, imaging, etc.) must be managed, stored, and analyzed. Many smaller projects (led by pharmaceutical companies, hospitals, and basic research groups) also seek to use the same analytical tools (and possibly data) to gain insight into specific aspects of different pathologies. The level of expectation is greater than at any point in medicine's recent history. However, this expectation will be frustrated if no integrative solutions are found for the huge computational demands generated by Personalized Medicine projects. There are unique issues associated with effectively managing genomic data for fundamental research and clinical practice, including the privacy of patient-related records as well as the computational infrastructure required to support the huge data volumes. Public data must be made available to allow effective meta-analyses of proprietary data. To achieve the increased statistical power needed to associate genetic variants with phenotypic traits, it must also be possible to share clinical records to find sufficient patients with specific characteristics. Overcoming these challenges would unleash the full potential of genomic data for research and clinical purposes.

The Human Technopole, Italy 2040 project seeks to develop a National Facility to provide Data Storage and High-Performance Computing (HPC) solutions to a community with huge data and CPU demands. This community faces complex restrictions in terms of security and ethical issues, limited technical HPC knowledge, and very severe time pressures in analyzing patient data. Personalized Medicine projects and services in the country's hospitals and wider biomedical research community will be supported by this National Facility in an efficient and sustainable manner. A reference site will be established to facilitate this community's access to validated standardized tools, datasets, and HPC resources. A single small initiative cannot tackle all these problems. Hence, nationwide action is needed to create a major data storage/HPC facility. This Facility will undertake joint endeavors in HPC and big data to create the basis for a national infrastructure. These endeavors include i) designing an integrated common sharing technology, ii) developing domain-oriented problem-solving solutions intended to become benchmarks and standards, iii) facilitating efficient access to databases of large volumes of structured and unstructured data, and iv) creating workflow strategies to close the dramatic gap between the next-generation informatics and computational tools and the ultimate goal of Personalized Medicine.

This aim will be achieved by creating mirror sites of publicly available genomic data and by enabling the electronic management of health records. To do this, it will be necessary to i) warehouse different large databases, ii) assimilate and process data from different sources, iii) provide domain-aware intelligent interfaces for discovering and accessing the data, and iv) process data with HPC and high-efficiency hardware. The Human Technopole is expected to generate petabytes-to-exabytes of data each year. To extract meaningful features and patterns, these data will need to be efficiently stored and promptly processed.

The Human Technopole and the Italian Supercomputer Center, CINECA, will collaborate to achieve these ambitious goals. A high-performance network connection will run between Bologna (the site of CINECA headquarter) and Milan in at the Expo site (where all the experimental activities and data generation will be performed). To transfer terabytes of data in a few minutes, the network infrastructure must carry at least 40 Gbit/s. The communication flow between Bologna and Milan will be increased up to a highly efficient 100 Gbit/s, which is the present level of network infrastructures worldwide.

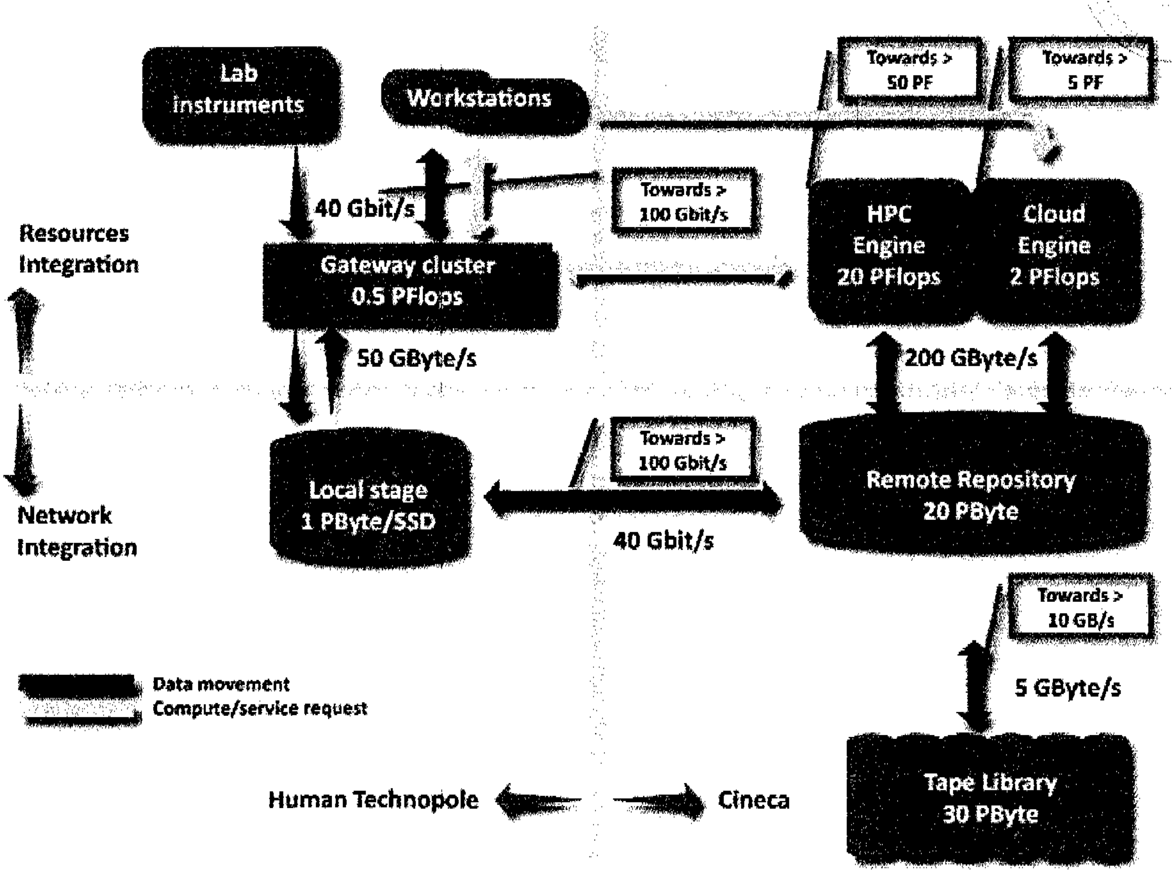


Figure 1. HPC and Data Storage infrastructure with hardware located at HT and CINECA.

HPC and Data Storage.

Big data storage facilities will be complemented by a large HPC infrastructure (i.e. machines with peak performance of tens of PFlops, see Figure 1). This combination of HCP with data storage is one of the Human Technopole's strategic assets. Until now, sequencing efforts have been too low-throughput in terms of the rate of genomes collected. As a result, HPC has not been deemed necessary. However, this will rapidly change. The nonlinear scaling nature of genome processing requires HPC capacity to grow more quickly than the data itself during data collection. The current computational paradigms are based mainly on single-server low-performing tools. Their analysis capacity decreases with respect to data volume, resulting in longer times for research efforts and decreased competitiveness. Given the limited efforts of the traditional bioinformatics community to optimize codes for HPC poses, it will be a unique challenge to create new-generation algorithms to concretely exploit HPC's full potential. Within the Human Technopole, computational and informatics activities will seek to rewrite existing algorithms (moving from interpreted to compiled languages) or, even better, to reformulate entire classes of problems. To do this, they will need to develop novel algorithms to tackle fundamental issues related to load balancing, synchronization, execution, and memory divergence. This will transform inherently inefficient codes into production-ready software (as an example, one possible strategy would be to implement codes on GPUs, Xeon Phis, Knights Landing, FPGAs, etc.). Image-based phenotypic screens are another area where big data storage and HPC will need to be combined. On the one side, big data storage will be needed to host and manage image files. On the other side, complex multiparametric image analyses will require HPC facilities so that meaningful information can be extracted within acceptable timeframes. HPC will also be useful in integrating and analyzing large heterogeneous databases from the socioeconomic domain, in collaboration with the CADS. Indeed, all the novel formalisms, algorithms, and codes developed within DSC, CLSC, and CADS will be fully optimized for exploitability and scalability in new-generation HPC hardware architectures. Finally, to achieve the ultimate goal of Personalized Medicine, a genomics approach must be combined with a molecular-based simulation approach. This will lead to a systems biology method, where the patient's characterized profile and pathology is combined with the design of personalized drugs and treatment. The capabilities enabled by an extreme computing performance will be fundamental to tackling this challenge.

The HPC and Data Storage Facility will be built in close collaboration with the Italian supercomputer center CINECA. CINECA recently installed a new computing system, named PICO, to respond to the ever-growing demand for services and capacity (storage, management, computing, and visualization). CINECA also has extensive experience in HPC infrastructures. Their newest machine, Marconi, whose first partition was installed in Q2 2016, will reach a peak performance of about 20 Pflops by the end of 2016. A team of 10-15 research technologists and technicians will be recruited internationally for the HPC and Data Storage Facility. They will develop specific workflows for data storage and for querying of next-generation petascale to exascale databases, together with a new user-friendly web portal for these activities. These novel tools will ultimately help experimentalists and medical doctors to more easily identify biologically and medically relevant patterns. The end user will thus benefit from a professional, flexible, and modern environment that facilitates the exploitation of HPC in Personalized Medicine.

Parallelization and Optimization of Codes and Software.

In strict collaboration with DSC, CLSC, and CADS, this Facility will establish a team of software engineers and computer scientists to develop new software and informatics tools to serve bioscience and life science. Formalisms, algorithms, and codes are designed and developed by theoretical and computational chemists, physicists, and bioinformaticians, who have profound knowledge of the scientific problem and the details and limits of the adopted models. However, there is a gap between this knowledge and the particular requirements of the newest computer architectures. This team's goal will be to bridge that gap. This is particularly important since a nonoptimal algorithmic implementation hampers the exploitation of the huge computational power becoming available in the HPC community. Moreover, a good knowledge of the underlying architecture is required to unlock the full potential of modern computer hardware. To mention two extremes, this modern computer hardware includes general purpose graphic-processing units (GP-GPUs) and multi-core microprocessors, including the next generation of system-on-socket microprocessors. It will thus be central to this Facility's mission to foster fruitful interactions between scientists and software engineers and to create a common language. This high-level support team will seek to achieve this goal by collaborating with and offering specialist support to users from the Human Technopole's entire scientific community.

An additional activity will be to create intelligent user interfaces, so that a wider community of users can access the power of the tools developed. In summary, this Facility will create optimized computational tools, which can then be transformed into new-generation software for computational chemistry and physics, bioinformatics, and big data analytics.

From the very start, algorithms and codes developed within the Human Technopole will be designed to be highly parallelizable and to run efficiently in HPC/big data environments (including CPU and GPU architectures). A team of software engineers and computer scientists will draw up community guidelines and standards to define how applications should be integrated (and if they need to be re-engineered and optimized) to develop solutions to specific Personalized Medicine problems. In particular, this Facility will anticipate what the new instrumentation, technologies, and methods will require in order to produce a fast response for the experimental activities. In addition, we will standardize and disseminate best practices for HPC/big data application actions. Personalized medicine is a young field. New techniques are generated, gain popularity, and are abandoned rapidly as more powerful new methods appear. One major task of this Facility will be to maintain an up-to-date record of the state-of-the-art technologies used both within the Human Technopole and at a national level.

In parallel, the Facility will create a list of software needed for the basic elements of a Personalized Medicine pipeline. These elements include analyzing genomic information; predicting the consequences of specific mutations and other genomic alterations, particularly within the exome; interpreting variation at the gene/protein network level; and identifying the potential drugs related to the predicted genetic alterations. These pipelines will also need interfaces so teams of clinicians, geneticists, genome experts, and bioinformaticians can interpret the results. These informatics tools will be routinely used by all Centers. They must be adapted or interfaced to other software modules so they can be included in complex workflows to meet the specific needs of different research activities and to fully exploit the HPC resources available. HPC will be a strategic and essential asset for the future of bioinformatics, big data analytics, and Personalized Medicine. In addition, HPC is increasingly needed for atomistic and multiscale simulations as the research becomes more ambitious and the biological systems investigated become more complex. Continuous effort is required to keep pace with rapidly evolving HPC architectures, which demand frequent modifications of programming paradigms. This often requires a substantial review of the code. In the last 20 years, there have been four disruptive paradigm changes: vector processing, massively parallel processing, shared-memory multicore processing, and GPU processing. If codes are not refactored in accordance with the new paradigms, they quickly become obsolete and can no longer be used on HPC architectures. The Facility's software libraries and datasets will be a valuable strategic asset. Their curation will require long-term planning and specialized resources. To achieve this, we will recruit a group of researchers and technologists. This group will be based at CINECA and will work for the Human Technopole's HPC/big data infrastructures.

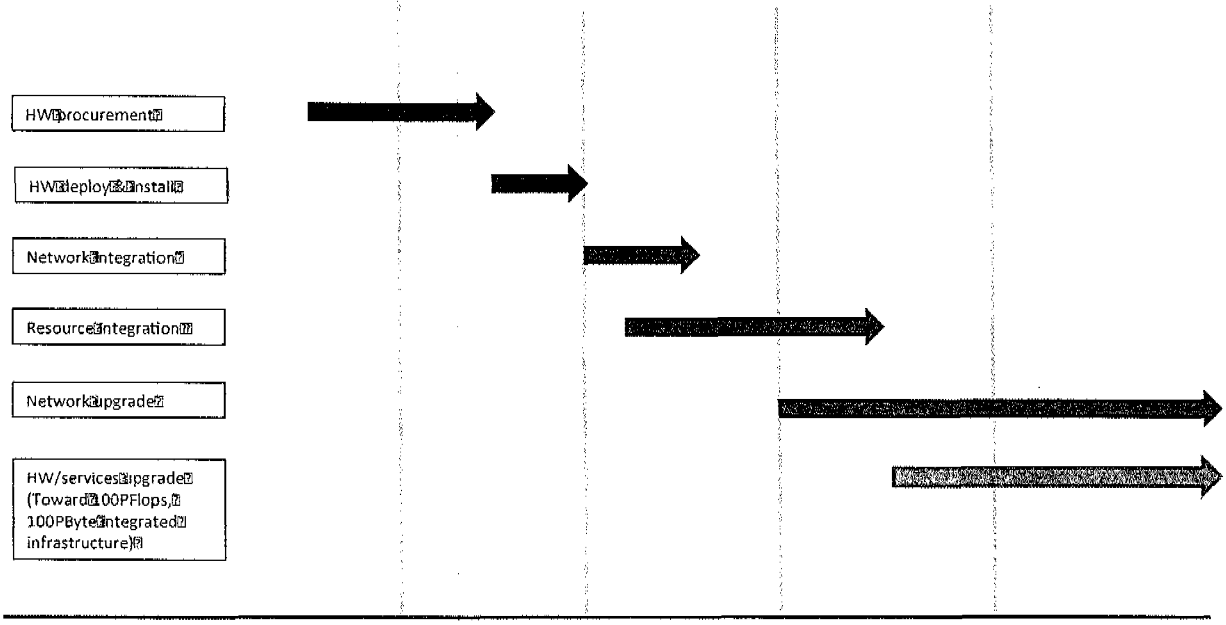


Figure 2. *Gantt for the creation of the F3 Facility for Data Storage and HPC and the activities involved.*

F

F4 FACILITY FOR COMMON SHARED SERVICES FOR IN VIVO AND IN VITRO MODELING

Vision: This facility will provide infrastructure and technologies for establishing, maintaining, and analyzing in vitro and in vivo models of disease.

Scientific Structure: F4 will comprise the iPSCs platform, together with the Mouse Genetics platform and Germ-Free Facility.

iPSCs Platform.

Vision. The iPSCs platform will function as a Core Facility for induced pluripotent stem cells (iPSCs), to integrate human whole-genome sequencing data with the establishment of iPSC lines from the same patients. The goal is to generate human neuronal subtypes to investigate pathophysiological mechanisms and carry out drug screening. iPSCs generated from human tumor samples will also be used to investigate tumorigenesis mechanisms and for different biomedical applications, including drug screening, toxicological testing, and biomarker selection. The platform will be responsible for iPSC generation, characterization, banking, and distribution to the wider Italian scientific community.

Mouse Genetics Platform and Germ-Free Facility.

Vision. This platform/facility will support the use of mice for in vivo studies of gene function and human-disease modelling, including the generation of compound mice and XPDs. Its services will include: *a)* Mouse husbandry (breeding, weaning, and maintenance; pathogen screening) and genotyping (tail biopsies and genotyping); *b)* Behavioral analysis and tracking (behavioral chambers, infrared cameras, tracking systems); *c)* Generation of genetically modified mice, using homologous recombination or genome-editing (TALENs & CRISPR) technologies; *d)* Rederivation and in vitro fertilization; *e)* Mouse sperm or embryo freezing; *f)* Germ-free facility (to grow mice in a completely sterile environment); and *g)* In vivo imaging (e.g. micro-CT, micro-PET, optical imaging, high-resolution ultrasound, two photon imaging).

PART 3 – FINANCIAL PLAN

1. INTRODUCTION

In this section we estimate the global financial needs of the Human Technopole (HT).

The financial needs of the Human Technopole vary in nature from the start-up to the steady state: refurbishment and adaptation of the buildings and laboratories set up will be the most important costs in the early two years, which will be partly sustained by the initial budget of 80 M€ allocated by the Government. OPEX and personnel costs in this period will be marginal.

After the start-up, staff and running costs are expected to grow continuously following the international recruitment, overcoming the Capital Expenditure at the third year. After year 4, the financial needs are determined primarily by the recruiting schedule (personnel and research budget of the scientists), and by the running costs of the laboratories. Obviously the final financial figures might change slightly depending on various external factors, such as the logistics of the EXPO area, the programs implemented by the PIs to be hired, and the recommendations of the future evaluation panels. However, at steady state the global financial needs can be estimated assuming a total headcount in the range of 1500 units, and **an average full cost per capita in the range of 95000 €** per year. This results in a total financial need in the range of 140 M€ per year at steady state. The full cost per capita is an average figure that includes the CAPEX investments (related to scientific equipment update and substantially reduced after the start-up phase) and the OPEX (labour and running costs) averaged per headcount.

Given the different characteristics of the research activities and the overall complexity of the HT masterplan, different levels of precision have been adopted for the financial estimates, namely:

- For the start-up (years 1 and 2) and the ramp up period (years 3 and 4) a very detailed estimate has been carried out, considering the foreseen buildup of the infrastructure, the hiring of the Principal Investigators (PIs) and of the assigned teams, assuming precise drivers and criteria for the budget allocation, as described in the following pages;
- Approaching steady state (i.e. from year 5 onwards), an approximate estimate has been made, extrapolating the growth of the centers up to the sixth year and assuming saturation with constant average full cost per capita afterwards.

Given these caveats, in what follows we discuss:

- *The Methodological Approach*, to highlight the criteria used in the estimate of the financial needs;
- *The Financial Estimates*, which classify the foreseen expenses by Center and by type of collaboration agreement; in the estimation we considered both Capital Expenses (CAPEX – i.e. durable goods and other investment in assets) and Operational Expenses (OPEX – i.e. utilities, consumables, services) and the cost of Human Resources;
- *The Time projections*, to describe the temporal evolution of the expenditure in the startup phase and at steady state.

2. METHODOLOGICAL APPROACH

a. Overview

The financial needs have been estimated according to the scheme depicted in Fig. 1, namely:

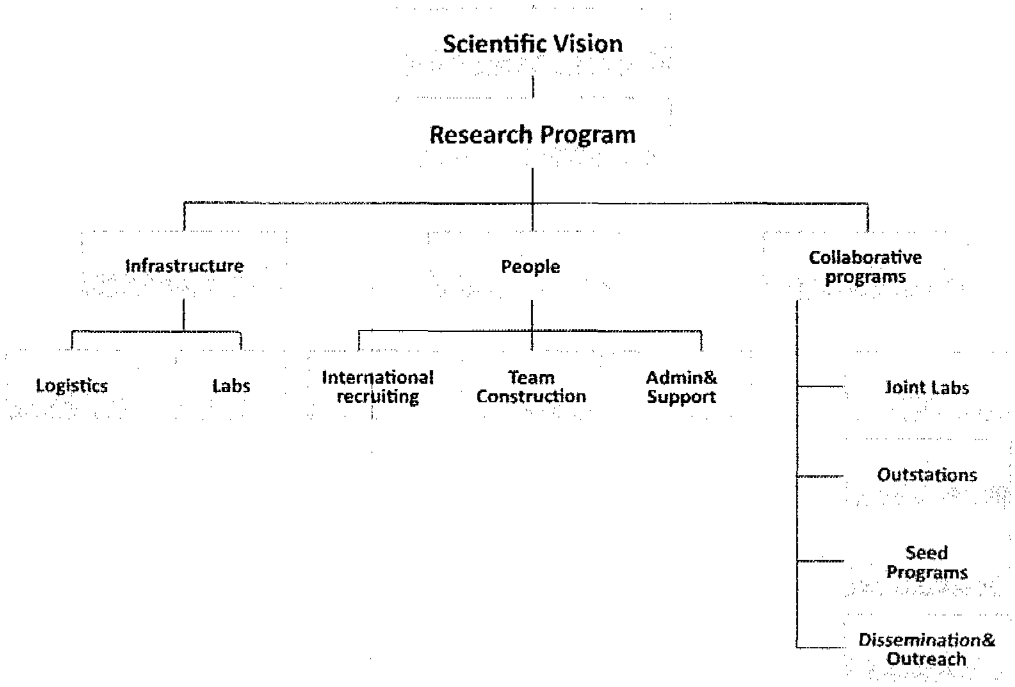


Fig. 1 - Financial Estimates Scheme

- the *Scientific Vision* of the Technopole results in a *Research Program*;
- The *Research Program* requires the creation of an *Infrastructure* consisting of Research Centers and Facilities;
- In each Center a number of *People* (Principal Investigators - PIs, research teams and support personnel) develop main stream Research Lines consisting of various Activities, also in collaboration with other Institutions by means of *Collaborative Programs* based on Joint Labs and Outstations.

b. Infrastructure

The infrastructure consists of two main items:

- Logistics: comprising all buildings, technical infrastructures and common spaces in which the Technopole will operate and accommodate labs and researchers; this includes all the ancillary infrastructure (e.g., power plant, technical gases piping and reservoirs, data communication lines, etc.) instrumental to the full operations of the research labs;
- Laboratories and Facilities: the actual scientific equipment and the equipped space in which research activities and experiments are carried out and where the research equipment is placed in operation.

The size of the infrastructure has been estimated taking into account:

- The average individual space *per capita* (in square meters - sqm) needed to accommodate people. According to IIT standards this amounts to: 12 sqm / person for computational people and 24 sqm / person for experimentalists;
- The common spaces (e.g. seminar rooms, services, warehouses, corridors etc.)
- Clearance, safety and optimal operation requirements of equipment intensive laboratories, following the design of each Center/Facility;
- The technical characteristics of the buildings, which determine the cost of the adaption, refurbishment, maintenance, utilities and other related expenses, and the running cost at steady state. Specifically for the refurbishment and adaptation of the building in the EXPO areas we took into account:
    - Official cost rates for construction and refurbishment works as recognized by the Lombardia Region;
    - Planning, design and project, around 10% of the total refurbishment cost;
    - Applicable VAT rates.

All the above parameters have been used to obtain the basic CAPEX needs as well as the basic OPEX for the operation of the logistics.

c. People

HT staff will consist of many research teams led by Principal Investigators (PIs). Each PI will lead a team of researchers and technologists, technicians, post docs, PhD students to develop the research activities and to run the equipment and instruments. In addition, the HT staff will include administration and support teams assisting the research teams, namely research organization, project administration, technology transfer, building and equipment maintenance, ICT, health & safety, administration and management, purchases and procurement etc. As discussed in Part 1 we expect up to 1500 people at regime to operate in the Technopole (including research staff and administrative and support staff, and excluding associated members from the Universities).

Scientific Teams

The estimate of the research headcounts has been made assuming standard teams to be formed by each PI. The recruiting of the research staff has also been projected in time, estimating an average of 24 months to complete the recruiting of a team. On average, a standard team consists of 6 units for a junior PI (tenure track entry level) and 12 units for a senior PI (tenured or senior tenure track). Team compositions may include staff researchers, technologists, post docs, technicians and PhD students (depending on research field) and determine the overall cost of personnel. In the calculations European level remunerations and average Italian scholarship costs for PhD students are assumed. HT Director and Centers' directors will be the first recruitments (first year of the project). The search for the other PIs will be initiated immediately afterwards. We expect approximately 40 PIs (including directors of the centers) to be hired in the first 2 years, out of about 100 PIs at steady state.

The foreseen hiring rate is highlighted in Table 1 and depicted by the plot in Fig. 2.

	Y1	Y2	Y3	Y4	Y5	Y6	Y7
Directors	7	7	7	7	7	7	7
PI	1	28	57	70	81	90	92
Scientific Teams	13	219	608	810	964	1.086	1.115
Admin. & Support	19	67	119	169	197	224	250
Grand Total	**40**	**321**	**791**	**1.056**	**1.249**	**1.407**	**1.464**

Table 1 – Cumulated staffing.

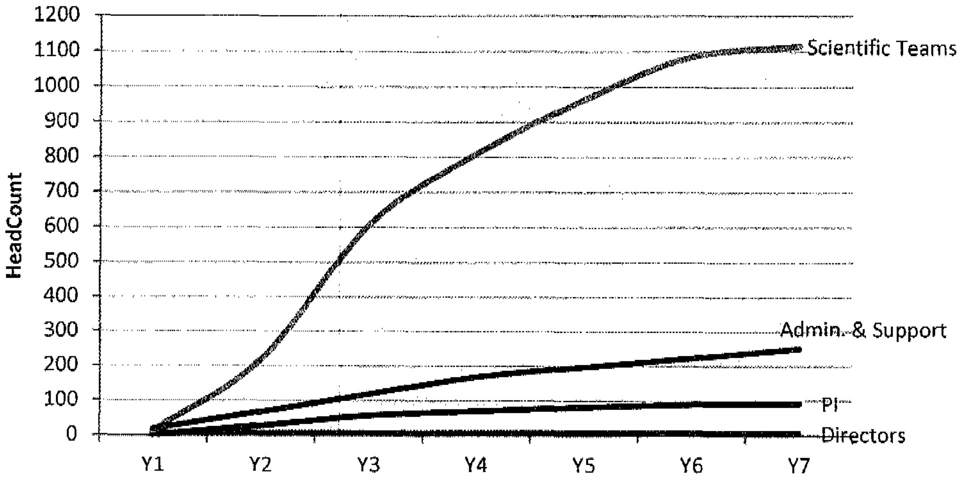


Fig. 2 - Cumulated staffing graph.

Administrative and Support Staff

Administrative and support staff is kept to the lowest possible level, in order to reduce as much as possible the global administrative overhead on research expenditure. According to IIT standards, the total administrative and research support cost (including technology transfer, patent office, technical office, research organization office, ICT, health & safety) is below 25% of the overall staffing.

d. Operational Expenses (OPEX)

OPEX for the research teams

Running costs of the research teams have been estimated using the following drivers: 16% of the average headcount cost for theoretical/computational activities and 40% of the average headcount cost for experimental activities. These OPEX include costs such as consumables (e.g. chemicals, biological materials, etc.), travels (e.g. participation in scientific conferences and seminars, cooperation with other research institutions on a worldwide basis), and services (e.g. lab equipment maintenance, publication fees, access to external databases, etc.).

OPEX related to support activities

Appropriate estimates have been conducted for other costs related to support activities (e.g. legal and statutory services, insurance, taxes, guard-services, etc.).

e. Other costs

Collaborative Programs

The complexity and multidisciplinary approach defined for the HT require the involvement of several Universities, Research Institutions and Research Hospitals; this involvement will have several forms, such as Joint Labs and Outstations (Outstations can be managed as research contracts or as IIT laboratories at the hosting Institutions). Joint Labs and Outstations will develop part of the research programs under the coordination of the Seven Centers of the HT.

Seed Programs

The HT could promote periodic calls for proposal (the Seed Programs), dedicated to external research institutions wanting to contribute to the development of the HT research program. The proposals evaluation should be done under the supervision of external panels (a Scientific and Technical Committee).

Cross-disciplinary programs

Collaborations and interactions among the Centers should be supported by periodic internal calls for cross-disciplinary programs to be developed jointly by teams belonging to different Centers, and evaluated by external panels (a Scientific and Technical Committee).

3. FINANCIAL NEEDS

Costs of the Centers

The following table highlights the estimated costs for each Center for the first 7 years. During the first 4 years the forecast is based on the drivers explained before. From the fifth year the forecast has been made considering the same standard costs for all centers.
A global HT assessment/evaluation should be performed after 4 years of activity, possibly affecting the forecast at steady state.

values in M€	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
Onco Genomics Center	9,7	7,7	15,4	15,3	16,2	19,3	19,8
Neuro Genomics Center	4,3	11,0	10,6	14,2	15,7	16,9	18,0
Agri-Food and Nutritional Genomics Center	6,6	8,0	11,3	12,8	14,7	16,3	17,0
Data Science Center	0,4	1,3	2,3	2,9	3,6	4,1	4,3
Center for Computational Life Sciences	5,8	2,9	5,7	7,5	8,5	9,3	10,1
Center for Analysis, Decisions, and Society	0,9	2,5	5,8	8,3	10,6	12,0	13,1
Center for Smart Materials and Devices	4,0	6,9	6,1	7,0	7,6	8,6	8,7
F1 - Central Genomics Facility	10,9	8,8	10,7	7,8	3,8	3,8	3,9
F2 - Imaging Facility	-	10,5	9,2	1,5	2,6	2,6	2,7
F3 - Data Storage & HPC Facility	0,3	5,6	5,6	5,6	6,7	6,8	6,8
F4 - Facility for Common Shared Services	0,0	4,7	6,4	1,6	2,6	2,7	2,8
Seed/Cross Disciplinary Programs	-	-	5,0	5,0	5,0	5,0	5,0
Research Lines Subtotal	**42,8**	**69,8**	**94,0**	**84,6**	**97,5**	**107,3**	**112,3**
Logistics	34,9	50,2	35,2	15,8	11,9	12,1	12,3
Admin. & Support	2,3	4,3	7,3	11,8	12,7	14,2	15,7
Total Human Technopole	**79,9**	**124,3**	**136,5**	**112,1**	**122,1**	**133,6**	**140,3**

Note:

- Each Center includes the costs of the HT-based labs and of the coordinated Outstations and Joint Labs

Collaborations

The following table provides a forecast of the cost of the external collaborations based on preliminary analysis of the activities to be carried out in the first 4 years (from the fourth year the estimation has been made considering standard costs applied to centers excluding collaboration agreements). Noticeably, the costs quoted in the table may vary depending on scientific plan, matching fund and human resources made available by the parties. In addition, scientific reviews that may affect the project can be scheduled after the second and the fifth year.

values in M€	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
Total Human Technopole	**79,9**	**124,3**	**136,5**	**112,1**	122,1	133,6	140,3
Human Technopole - IIT	**72,8**	**104,4**	**107,6**	**79,5**			
HT-Research	35,6	50,0	65,1	52,0			
HT- Infrastructure/Campus	8,5	21,8	20,5	17,0			
HT - Building refurbishment	26,6	28,9	15,5				
Admin & Support	2,0	3,9	6,5	10,6			
Joint Labs	**5,5**	**8,2**	**13,7**	**14,2**			
UniMi	5,4	3,0	7,0	6,0			
Bicocca	0,1	2,9	2,6	3,0			
PoliMi	0	2,3	4,1	5,1			
IIT @ OutStations	**1,3**	**8,0**	**9,7**	**12,0**			
IIT @ INGM	0,2	0,6	1,2	1,7			
IIT @ OSR	0,1	0,7	2,2	2,2			
IIT @ Humanitas	0,3	3,7	1,7	2,7			
IIT @ BESTA	0,3	2,0	2,0	2,0			
IIT @ Mario Negri	0,4	1,1	1,2	1,2			
IIT @ FEM	0	0,0	0,9	1,5			
IIT @ Cineca	0	0,0	0,5	0,7			
Research Contract	**0,4**	**3,7**	**5,5**	**6,5**			
IEO	0,0	0,5	1,1	1,1			
INT	0,0	0,6	0,9	1,0			
PoliMi	0,0	0,2	0,2	0,3			
CREA	0,0	0,9	0,9	0,9			
Parco Tecnologico Padano	0,0	0,2	0,2	0,2			
ISI	0,4	1,3	2,3	2,9			

Estimation done according to headcount development and standard costs

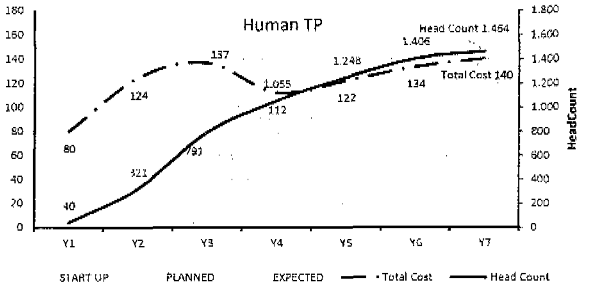
4. TIME PROJECTIONS

The following tables and graphs depict the time evolution of the foreseen expenses (values in M€) by area, namely:

- Total, for the overall expenses;
- Research Area, for the expenses of the research centers, facilities and outstations;
- Logistics and set-up, for the expenses needed to set up the HT;
- Administration and Support, for the activities related to the support of the research centers and the overall management of the initiative.

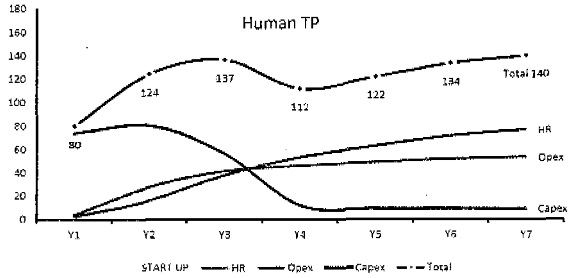
In the definition of the time projection of the expenses, appropriate synchronicity has been assumed between research teams construction and realization of the infrastructure, as defined in section 1.

Scientific reviews that may affect the project can be scheduled after the second and the fifth year.



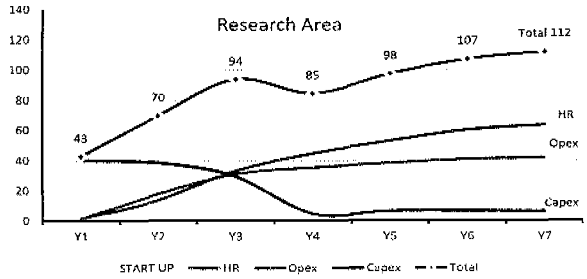
Total

Year	Y1	Y2	Y3	Y4	Y5	Y6	Y7
Total Cost	79,9	124,3	136,5	112,1	122,1	133,6	140,3
Head Count	40	321	791	1.055	1.248	1.406	1.464
Per capita in K€			173	106	98	95	96



Total

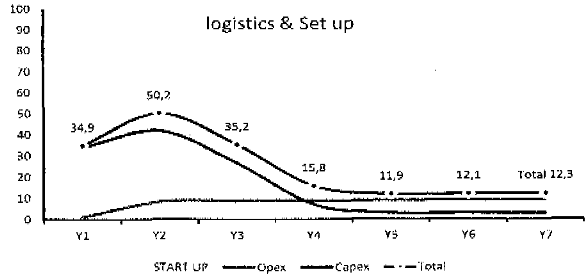
Year	Y1	Y2	Y3	Y4	Y5	Y6	Y7
HR	2,4	16,0	38,2	53,5	63,4	72,0	77,1
Opex	3,5	27,7	41,8	46,2	49,3	52,2	53,7
Capex	74,1	80,6	56,6	12,3	9,4	9,4	9,4
	79,9	124,3	136,5	112,1	122,1	133,6	140,3



Research Area

Total

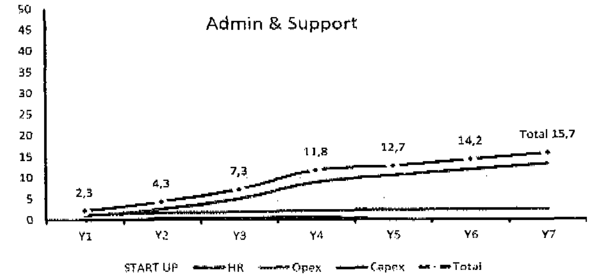
Year	Y1	Y2	Y3	Y4	Y5	Y6	Y7
HR	1	14	33	45	53	60	64
Opex	1	18	31	35	38	41	42
Capex	40	39	30	5	6	6	6
	42,8	69,8	94,0	84,6	97,5	107,3	112,3



Logistics & Setup

Total

Year	Y1	Y2	Y3	Y4	Y5	Y6	Y7
HR	0	0	0	0	0	0	0
Opex	1	8	9	9	9	9	9
Capex	34	42	27	7	3	3	3
	34,9	50,2	35,2	15,8	11,9	12,1	12,3



Admin & Support

Total

Year	Y1	Y2	Y3	Y4	Y5	Y6	Y7
HR	1	2	5	9	10	12	13
Opex	1	2	2	2	2	2	3
Capex	0	0	0	1	0	0	0
	2,3	4,3	7,3	11,8	12,7	14,2	15,7

Fig.3 Summary plots of the financial needs of the Human Technopole (in M €).

5. CENTERS DETAIL

In the following pages we exhibit the financial and staffing estimates for the various centers of the Human Technopole, providing for each a brief rationale.
For each of the centers we provide:

- Two tables that show the evolution of the overall spending for the center and the cumulative hiring, and the subdivision of the spending between internal HT and other locations;
- Three graphs that illustrate respectively the trend of the spending subdivision between internal HT and other location; the evolution of the staffing and the overall spending; the evolution of the spending by nature.

F

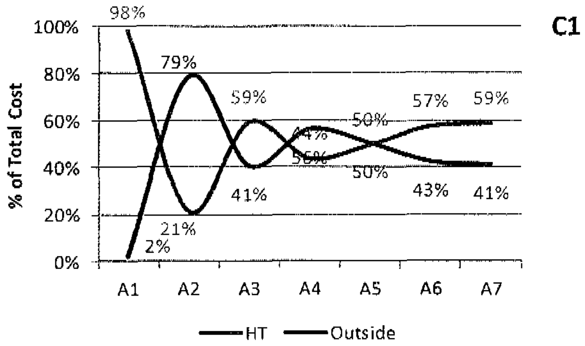
a. C1- Onco Genomics Center

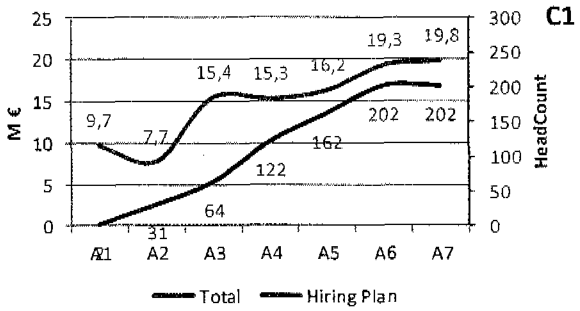
C1- Onco Genomics Center							
M€	A1	A2	A3	A4	A5	A6	A7
Human Resources	0,2	1,5	4,0	7,0	8,3	10,5	10,9
Opex	0,1	2,4	4,7	6,9	7,4	8,3	8,4
Capex	9,4	3,9	6,8	1,4	0,5	0,5	0,5
Total	9,7	7,7	15,4	15,3	16,2	19,3	19,8
Hiring Plan	2	31	64	122	162	202	202

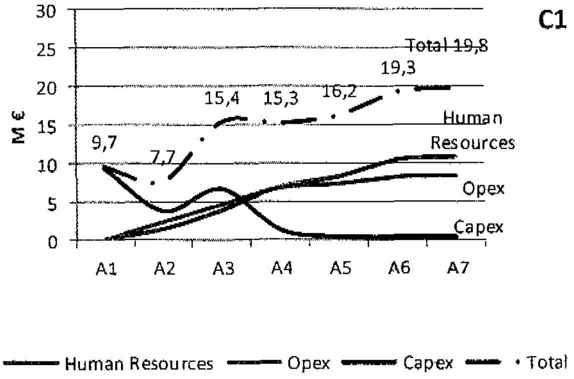
M€	A1	A2	A3	A4	A5	A6	A7
HT	9,5	1,7	9,1	6,7	8,1	11,1	11,6
Outside	0,2	6,1	6,3	8,6	8,2	8,2	8,2
Total	9,7	7,7	15,4	15,3	16,2	19,3	19,8

Center C1 will reach steady state with approximately 200 staff and 19,8 M€/year, and 41% of the activities in outstations (primarily Research Hospitals).
The laboratories' infrastructure will be completed in about 3 years. The full cost per capita at regime will be in the range of 98 K€/year.

per capita cost in K€	98







7

b. C2 - Neuro Genomics Center

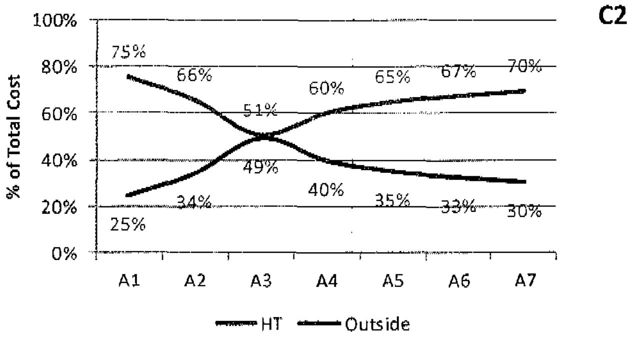
C2 - Neuro Genomics Center							
M€	A1	A2	A3	A4	A5	A6	A7
Human Resources	0,2	2,0	5,2	7,9	9,4	10,2	11,0
Opex	0,8	3,1	4,2	5,3	5,8	6,2	6,5
Capex	3,3	5,9	1,2	1,1	0,5	0,5	0,5
Total	4,3	11,0	10,6	14,2	15,7	16,9	18,0
Hiring Plan	2	34	108	152	177	192	197

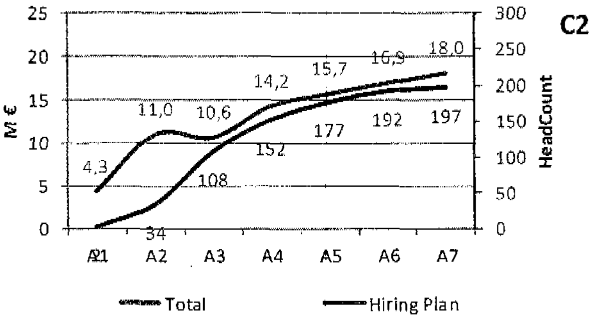
M€	A1	A2	A3	A4	A5	A6	A7
HT	3,2	7,2	5,4	8,6	10,2	11,4	12,5
Outside	1,1	3,8	5,2	5,7	5,5	5,5	5,5
Total	4,3	11,0	10,6	14,2	15,7	16,9	18,0

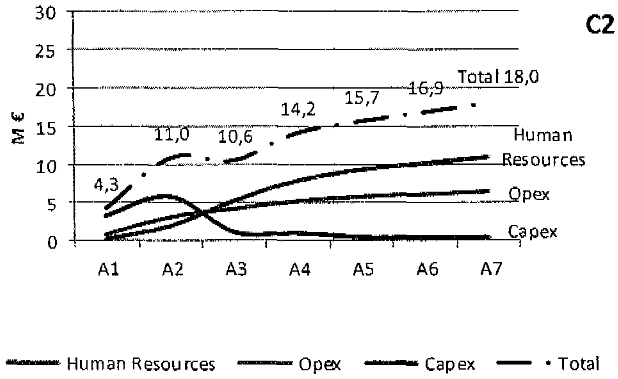
per capita cost in K€	92

Center C2 will reach steady state with approximately 197 staff and 18 M€/year, and 30% of the activities in outstations (primarily Research Hospitals).

The laboratories' infrastructure will be completed in about 3 years. The full cost per capita at regime will be in the range of 92 K€/year.







F

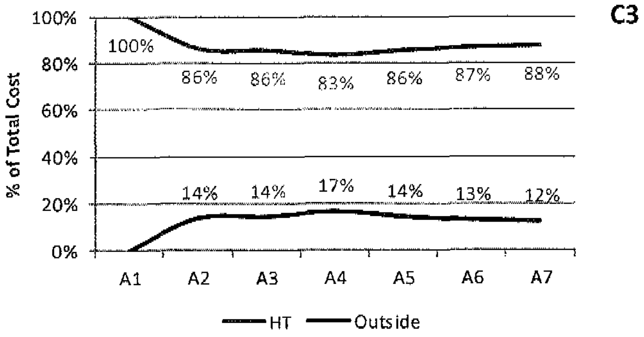
c. C3 - Agrifood & Nutritional

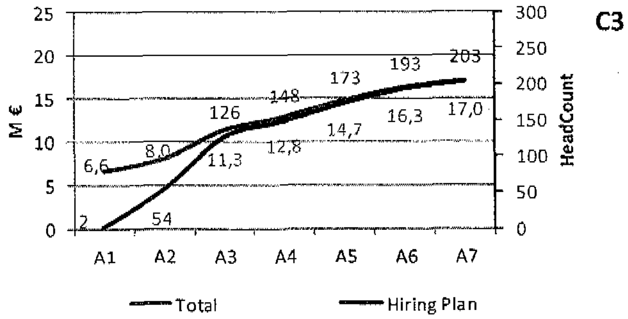
C3 - Agrifood & Nutritional							
M€	A1	A2	A3	A4	A5	A6	A7
Human Resources	0,2	3,1	6,9	8,1	9,5	10,6	11,1
Opex	0,1	2,3	3,8	4,2	4,8	5,2	5,4
Capex	6,4	2,6	0,6	0,5	0,5	0,5	0,5
Total	6,6	8,0	11,3	12,8	14,7	16,3	17,0
Hiring Plan	**2**	**54**	**126**	**148**	**173**	**193**	**203**

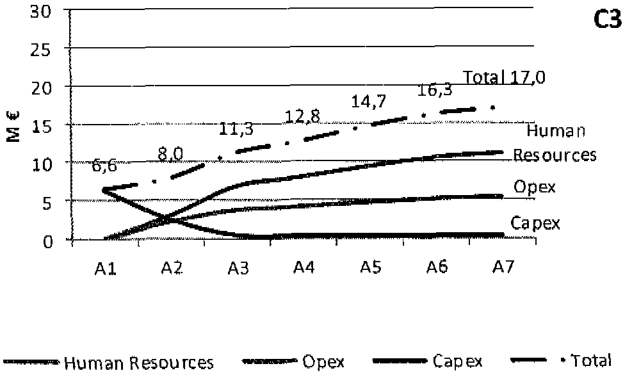
M€	A1	A2	A3	A4	A5	A6	A7
HT	6,6	6,9	9,7	10,7	12,6	14,1	14,9
Outside	0,0	1,1	1,6	2,1	2,1	2,1	2,1
Total	6,6	8,0	11,3	12,8	14,7	16,3	17,0

Center C3 will reach steady state with approximately 203 staff and 17 M€/year, and 12% of the activities in outstations (primarily Research Labs). The laboratories' infrastructure will be completed in about 2 years. The full cost per capita at regime will be in the range of 84 K€/year.

per capita cost in K€	84







d. C4 - Data Science Center

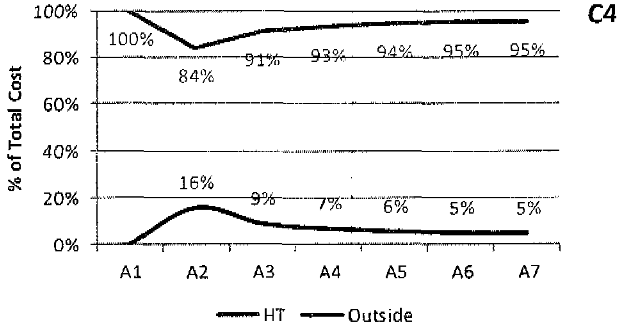
C4 - Data Science Center							
M€	A1	A2	A3	A4	A5	A6	A7
Human Resources	0,2	0,8	1,7	2,3	2,8	3,3	3,4
Opex	0,0	0,3	0,5	0,6	0,7	0,7	0,7
Capex	0,2	0,1	0,1	0,1	0,1	0,1	0,1
Total	0,4	1,3	2,3	2,9	3,6	4,1	4,3
Hiring Plan	**2**	**19**	**31**	**36**	**44**	**50**	**50**

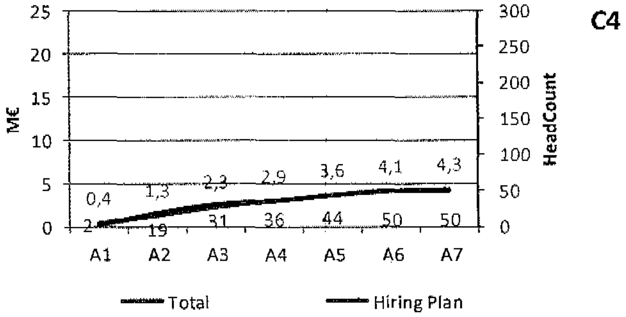
M€	A1	A2	A3	A4	A5	A6	A7
HT	0,4	1,1	2,1	2,7	3,4	3,9	4,1
Outside	0,0	0,2	0,2	0,2	0,2	0,2	0,2
Total	0,4	1,3	2,3	2,9	3,6	4,1	4,3

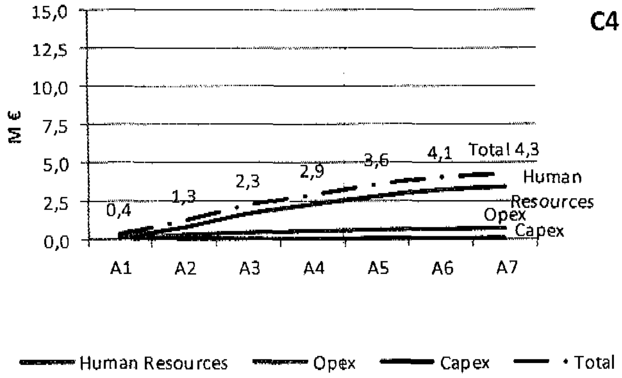
per capita cost in K€	86

Center C4 will reach steady state with approximately 50 staff and 4,3 M€/year, and 5% of the activities in outstations (primarily Research Labs).

The computational infrastructure will be completed in 1 year. The full cost per capita at regime will be in the range of 86 K€/year.







F

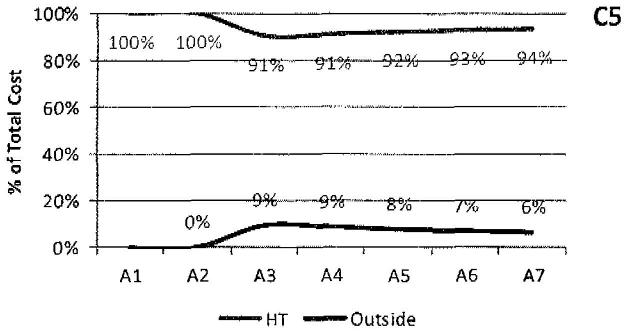
e. C5 - Computational Life Sciences

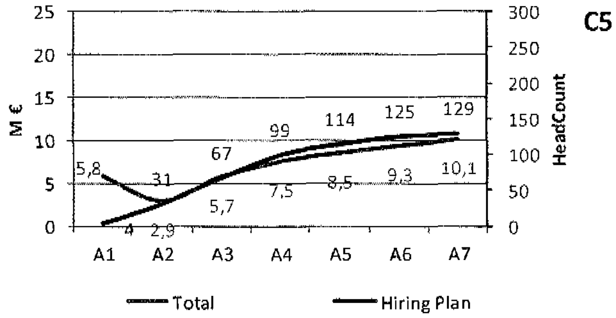
C5 - Computational Life Sciences							
M€	A1	A2	A3	A4	A5	A6	A7
Human Resources	0,2	1,8	3,9	5,3	6,1	6,7	7,2
Opex	0,1	0,7	1,4	1,8	2,2	2,4	2,6
Capex	5,5	0,4	0,4	0,4	0,2	0,2	0,2
Total	5,8	2,9	5,7	7,5	8,5	9,3	10,1
Hiring Plan	4	31	67	99	114	125	129

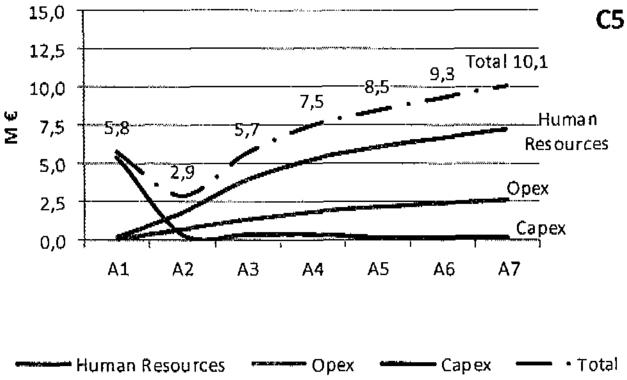
M€	A1	A2	A3	A4	A5	A6	A7
HT	5,8	2,9	5,2	6,8	7,8	8,7	9,4
Outside	0,0	0,0	0,5	0,7	0,7	0,7	0,7
Total	5,8	2,9	5,7	7,5	8,5	9,3	10,1

Center C5 will reach steady state with approximately 129 staff and 10.1 M€/year, and 6% of the activities in outstations (primarily Research Labs). The computational infrastructure will be completed in 2 years. The full cost per capita at regime will be in the range of 78 K€/year.

per capita cost in K€	78







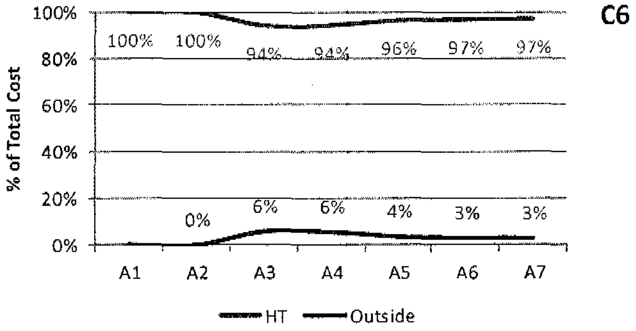
f. C6 - Center for Analysis, Decisions & Society

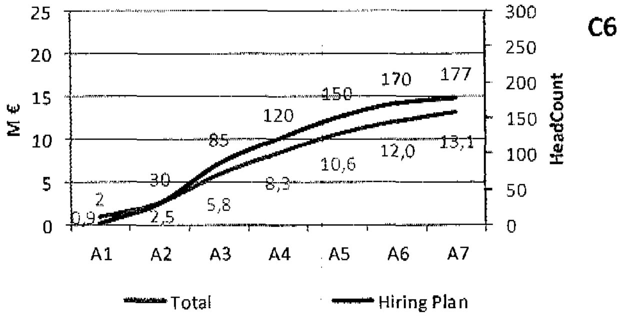
C6 - Center for Analysis, Decisions & Society							
M€	A1	A2	A3	A4	A5	A6	A7
Human Resources	0,2	1,8	4,5	6,4	8,1	9,3	10,1
Opex	0,0	0,5	1,3	1,9	2,3	2,7	2,9
Capex	0,7	0,2	0,0	0,0	0,1	0,1	0,1
Total	0,9	2,5	5,8	8,3	10,6	12,0	13,1
Hiring Plan	2	30	85	120	150	170	177

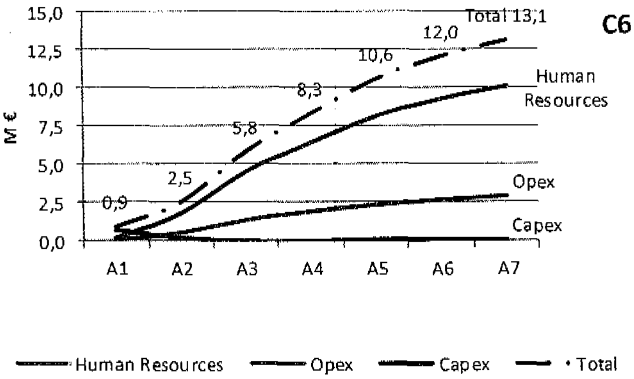
M€	A1	A2	A3	A4	A5	A6	A7
HT	0,9	2,5	5,5	7,8	10,2	11,7	12,7
Outside	0,0	0,0	0,3	0,5	0,4	0,4	0,4
Total	0,9	2,5	5,8	8,3	10,6	12,0	13,1

per capita cost in K€	74

Center C6 will reach steady state with approximately 177 staff and 13.1 M€/year. The center will be joint (50% matching fund) with Politecnico di Milano with little (3%) activity in outstations. The software/hardware infrastructure will be completed in 2 years. The full cost per capita at regime will be in the range of 74 K€/year.







g. C7 - Center for Smart Materials and Devices

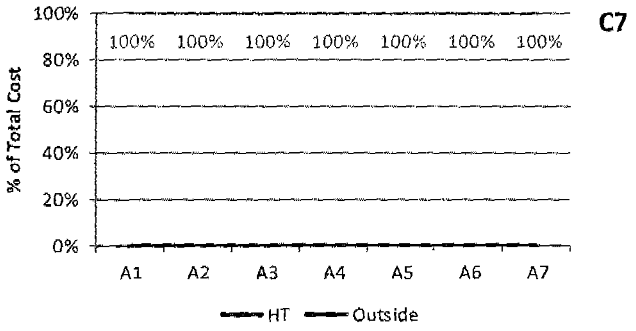
C7 - Center for Smart Materials and Devices							
M€	A1	A2	A3	A4	A5	A6	A7
Human Resources	0,2	1,3	3,6	4,3	5,2	5,8	5,9
Opex	0,1	0,6	1,4	1,7	2,0	2,2	2,3
Capex	3,8	5,0	1,0	1,0	0,5	0,5	0,5
Total	4,0	6,9	6,1	7,0	7,6	8,6	8,7
Hiring Plan	2	23	77	95	113	128	130

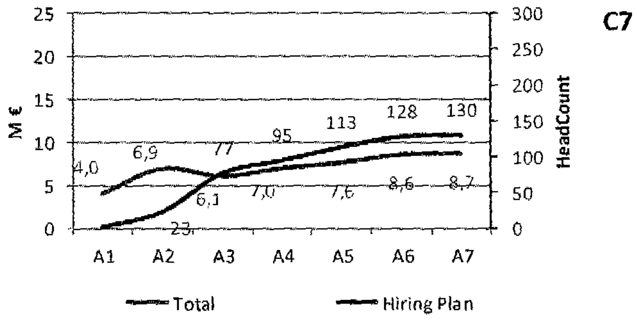
M€	A1	A2	A3	A4	A5	A6	A7
HT	4,0	6,9	6,1	7,0	7,6	8,6	8,7
Outside	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Total	4,0	6,9	6,1	7,0	7,6	8,6	8,7

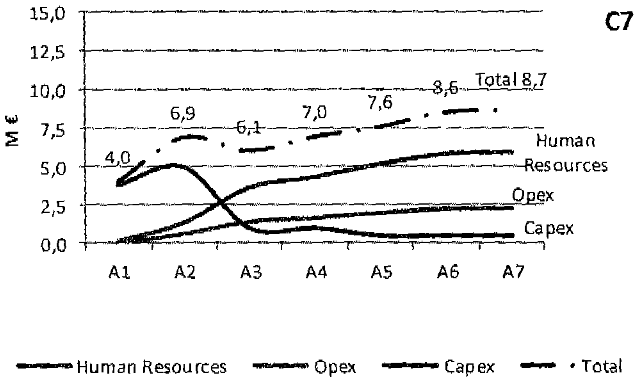
per capita cost in K€	67

Center C7 will reach steady state with approximately 130 staff and 8,7 M€/year. The center will exploit the nanotech infrastructure of the Istituto Italiano di Tecnologia in Milan (IIT Center for Nano Science and Technology located at Politecnico di Milano).

The laboratories' infrastructure will be completed in 3 years. The full cost per capita at regime will be in the range of 67 K€/year.





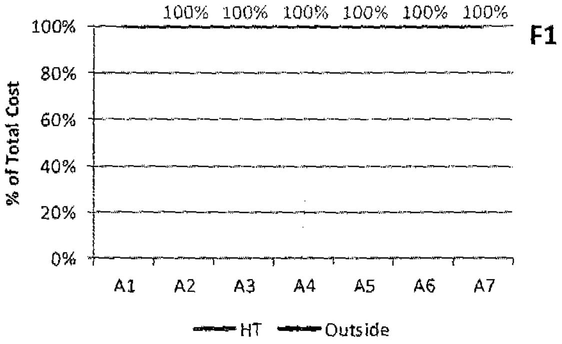


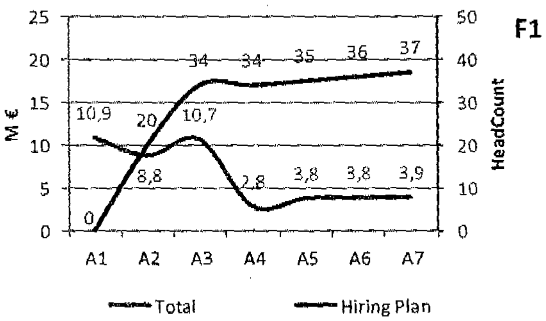
h. F1 – Central Genomics Facility

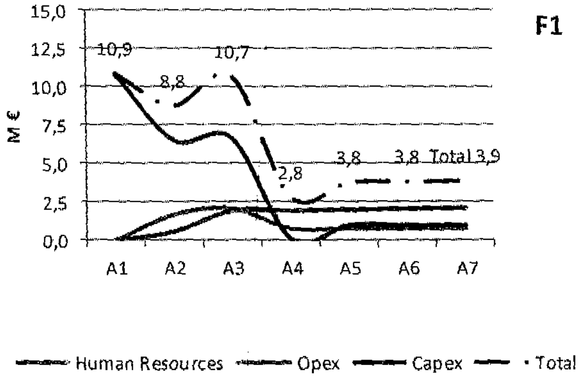
Facility F1 will reach steady state with approximately 37 staff and 3,9 M€/year. The laboratories' infrastructure will be completed in 4 years.

F1 - Central Genomics Facility							
M€	A1	A2	A3	A4	A5	A6	A7
Human Resources	0,0	0,6	1,9	1,9	2,0	2,0	2,1
Opex	0,0	1,7	2,1	0,8	0,8	0,8	0,8
Capex	10,9	6,5	6,7	0,1	1,0	1,0	1,0
Total	10,9	8,8	10,7	2,8	3,8	3,8	3,9
Hiring Plan	**0**	**20**	**34**	**34**	**35**	**36**	**37**

M€	A1	A2	A3	A4	A5	A6	A7
HT	10,9	8,8	10,7	2,8	3,8	3,8	3,9
Outside	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Total	10,9	8,8	10,7	2,8	3,8	3,8	3,9





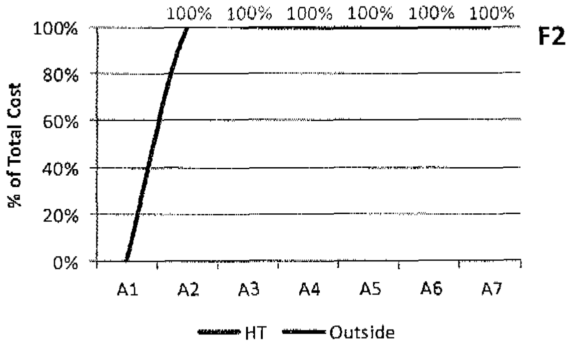


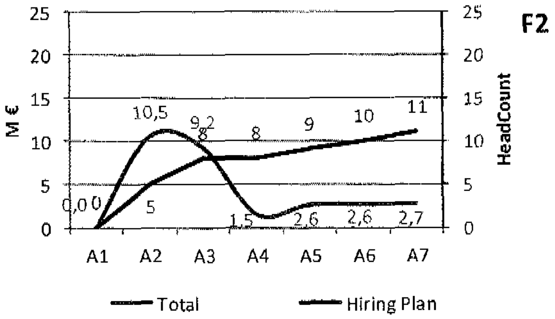
i. F2 – Imaging Facility

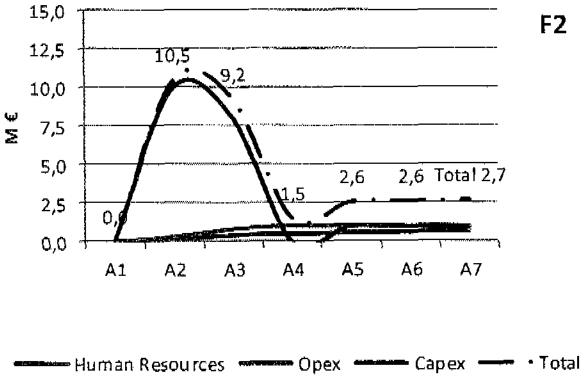
F2 - Imaging Facility							
M€	A1	A2	A3	A4	A5	A6	A7
Human Resources	0,0	0,2	0,4	0,5	0,6	0,6	0,7
Opex	0,0	0,3	0,8	1,0	1,0	1,0	1,0
Capex	0,0	10,0	8,0	0,0	1,0	1,0	1,0
Total	0,0	10,5	9,2	1,5	2,6	2,6	2,7
Hiring Plan	**0**	**5**	**8**	**8**	**9**	**10**	**11**

M€	A1	A2	A3	A4	A5	A6	A7
HT	0,0	10,5	9,2	1,5	2,6	2,6	2,7
Outside	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Total	0,0	10,5	9,2	1,5	2,6	2,6	2,7

Facility F2 will reach steady state with approximately 11 staff and 2,7 M€/year. The laboratories' infrastructure will be completed in 4 years.







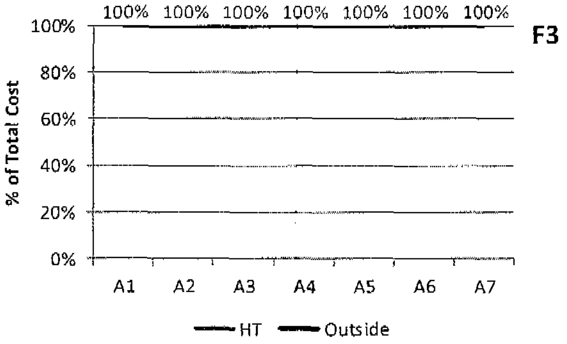
j. F3 – Data Storage and High Performance Computing Facility

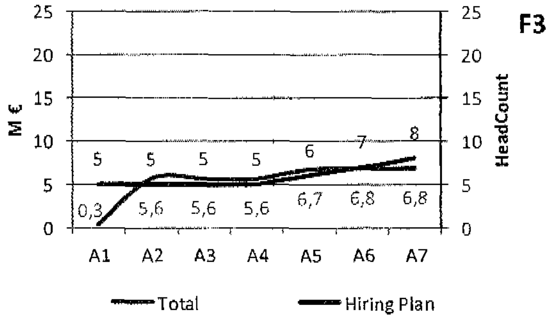
Facility F3 will be located at Cineca (100% outstation). It will reach steady state with approximately 8 staff and 6,8 M€/year. The laboratories' infrastructure will be updated constantly.

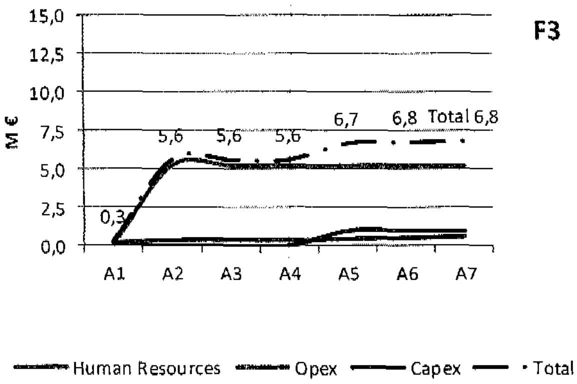
F3 - Data Storage & HPC Facility

M€	A1	A2	A3	A4	A5	A6	A7
Human Resources	0,2	0,4	0,4	0,4	0,5	0,5	0,6
Opex	0,1	5,2	5,2	5,2	5,2	5,2	5,2
Capex	0,0	0,0	0,0	0,0	1,0	1,0	1,0
Total	0,3	5,6	5,6	5,6	6,7	6,8	6,8
Hiring Plan	**5**	**5**	**5**	**5**	**6**	**7**	**8**

M€	A1	A2	A3	A4	A5	A6	A7
HT	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Outside	0,3	5,6	5,6	5,6	6,7	6,8	6,8
Total	0,3	5,6	5,6	5,6	6,7	6,8	6,8







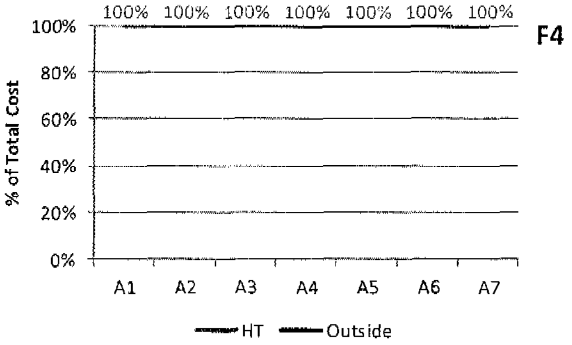
k. F4 –Facility for Common Shared Services

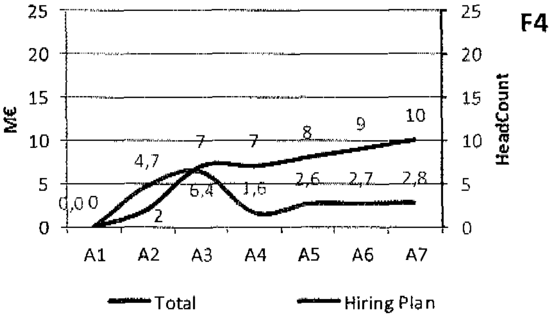
Facility F4 will reach steady state with approximately 10 staff and 2,8 M€/year. The laboratories' infrastructure will be completed in 4 years.

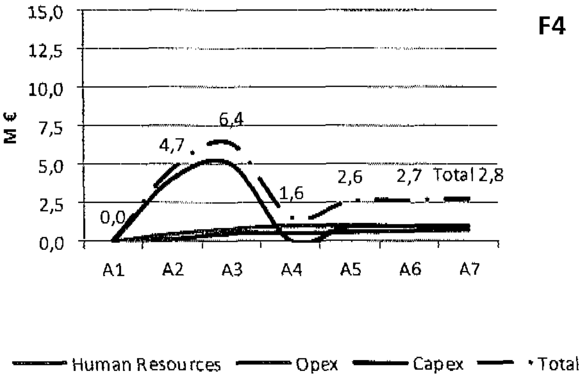
F4 - Facility for Common Shared Services

M€	A1	A2	A3	A4	A5	A6	A7
Human Resources	0,0	0,1	0,5	0,5	0,6	0,7	0,8
Opex	0,0	0,5	0,8	1,0	1,0	1,0	1,0
Capex	0,0	4,1	5,1	0,1	1,0	1,0	1,0
Total	0,0	4,7	6,4	1,6	2,6	2,7	2,8
Hiring Plan	**0**	**2**	**7**	**7**	**8**	**9**	**10**

M€	A1	A2	A3	A4	A5	A6	A7
HT	0,0	4,7	6,4	1,6	2,6	2,7	2,8
Outside	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Total	0,0	4,7	6,4	1,6	2,6	2,7	2,8







APPENDIX 1

Endorsement letters:

- Camera di Commercio di Milano
- INAIL - Istituto Nazionale per L'Assicurazione contro gli Infortuni sul Lavoro
- IBM Research GmbH
- Compagnia di San Paolo
- Fondazione Don Carlo Gnocchi
- Regione Emilia-Romagna
- Fondazione Umberto Veronesi
- Orogel fresco
- Regione Umbria
- Labomar
- Consorzio Innovazione Frutta
- Assosementi - Associazione Italiana Sementi
- Assomela - Associazione Italiana Produttori di Mele
- Apofruit
- Apoconerpo
- Winegraft
- CIVIT – Consorzio Innovazione Vite
- Wine Research Team

F

APPENDIX 2

International recruitment procedure (see also https://www.iit.it/careers/working-iit)

HT will recruit technical and scientific staff focusing on the quality of candidates and adopting the selection procedures of IIT.

IIT has established a Standing Committee of External Evaluators (SCEE), a list of particularly qualified scientists at international level in the fields covered by the strategic plan of the Institute, who take part in the selection and evaluation of the scientific staff. The activity of these external members is performed pro bono.

Tenure Track

Principal Investigators (PIs) at IIT are tenure track or tenured scientists. They carry out their own research activity (see research lines) in the frame of the Strategic Plan, and they are fully independent.

Tenure track scientists are hired upon international calls. An international Search Committee is established by the Scientific Technical Committee (STC) to evaluate applications and create a short list of candidates to interview. Short listed candidates are invited for a public presentation and an interview in front of an Evaluation Committee consisting of a majority of external experts. Candidates may access the tenure track at junior (stage 1) or senior level (stage 2). The maximum duration of the tenure track is 5 + 5 years (stage 1 + stage 2). A peer review evaluation must be passed at the end of each phase. Senior scientists can also be recruited as tenured scientists upon evaluation of the Scientific Technical Committee.

Staff Researchers

Researchers are hired by open calls (scientific journals and IIT web site), through Search and Evaluation panels consisting of both internal and external experts. Researchers positions are time limited (up to five years). Researchers report to a PI and can be in charge of a laboratory and post docs and students.

Technologists

Technologists are hired by open calls (scientific journals and IIT web site), through Search and Evaluation panels consisting of both internal and external experts. Their primary mission is to develop technology-oriented research and/or coordinate laboratories and facilities of general interest for the Institute. Technologists report to a PI and can be in charge of post docs, students and technicians.

Post docs

Post docs are recruited by open calls (scientific journals and IIT web site), through Search and Evaluation panels consisting of internal experts. The maximum duration of a Post doc appointment at IIT is 5/6 years. Post docs are considered junior if recruited within 2 or 3 years after their PhD diploma, whereas the senior level is for those who have at least 3 years' experience after the PhD.

Technical and Administrative Staff

The recruiting of technical and administrative staff is established on a competitive basis, through publication of advertisements on the IIT website and specialized websites for recruiting. The advertisement defines the position to be filled, the main activities, the technical, behavioral and managerial skills requested for the role, methods and timing for the application. The evaluation committee, composed by

F

the specialists of the Human Resources and Organization Directorate and by the head of the Organizational Unit/Research Line, screens all the applications and select a list of candidates. The skills and the level of the experience required are evaluated by individual or group interviews and technical tests. For some professional profiles, the selection process may include a first step consisting of an Assessment Center, a methodology that help assessors identify how close candidates' abilities and behaviors match the sought-after job profile.

The evaluation committee defines a final ranking and identifies the successful candidate following the outcome of the interviews and the tests.

APPENDIX 3

EXPECTED ECONOMIC IMPACT, LONG TERM SUSTAINABILITY AND POTENTIAL REVENUES OF THE HUMAN TECHNOPOLE

The expected economic impact of HT should be evaluated on a time scale of at least 10 years. The development of:

- a precision medicine and preventive nutrition strategy for public health,
- new technologies for food traceability, production and packaging,
- new medical diagnostics approaches and the development of new software,
- new algorithms and predictive models,

will impact remarkably at technological as well as social and economical level. There are three main indicators of impact to be considered:

1- direct fund raising of the research;
2- creation of new jobs and entrepreneurial activities;
3- reduction of public expenditure induced by the HT technologies on the public health system and other public domains

Direct fund raising will be a primary target of HT. Competitive fund raising channels include: European programs (eg Horizon 2020 and ERC), other national and international individual awards, including charity funds, industrial grants and sponsored research agreements. We also expect patent licensing and IP related revenues (eg royalties) to contribute to the global fund raising of HT. **At steady state HT targets to raise yearly up to 40% of the research cost.**

In addition to fund raising, a positive impact of HT on GNP is expected by the creation of new startup and new jobs and by the transfer of technologies to companies and hospitals. Enhanced attractiveness of international companies in Italy and establishment of joint public/private R&D laboratories between HT and industries should also be considered. Despite these actions do not impact directly on the finances of HT, they represent positive indicators of impact on the country's economy, which is ultimately measured by the number of new jobs created around the Human Technopole activities. **A good target could be to double the number of jobs around the HT initiative at steady state**.

The most important indicator is obviously the expenditure reduction induced by the massive impact of the HT program on the public health system, food and nutrition system, predictive models for social needs and decision making system in the long term. It is very difficult to make quantitative prediction at this stage. However, considering that the social cost of Cancer and Neurodegenerative diseases in Italy amounts to about 2% of the GNP (> 30 Billion Euro) per year, even a partial success of the precision medicine strategy would by far compensate the investment of the first 10 years of HT. Similar arguments apply to the food/nutrition pipeline and to the predictive models applied to other social domains (such as the tax system).

A more detailed analysis of expected impact of the different HT centers is provided in the following.

CENTER C1 and C2: IMPACT ON THE PUBLIC HEALTH AND PUBLIC EXPENDITURE

Biopharma and care executives are optimistic that Personalised Medicine will improve efficacy, safety and public health (through improved disease prevention, management of disease in early stages, prediction of a

specific patient's clinical response to various medical interventions). It is also generally perceived that Personalized Medicine may be economically viable, due to: i) improved primary and preventive care (expected to greatly reduces future health care costs) and ii) improved disease control (expected to reduce global costs of current disease treatments by increasing efficacy).
There is, however, a widespread skepticism about the financial impact of Personalized Medicine in the short term. In particular, since Personalized Medicine is based on innovation through science and technology, it is automatically suspected to be unsustainable in terms of spending (advances in medical technology are widely thought to contribute significantly to the escalating costs of health care). This is mainly due, however, to the fact that there are only a few studies available that analyzed Personalized-Medicine interventions (economic evaluations are largely based on drugs designed to treat the whole patient population). Accurate measurements of the economical effects of Personalized Medicine is indeed among the top priorities of HT. The few available studies, however, demonstrate the potential of Personalized Medicine to reduce costs of Health.
Measurements of cost-effectiveness or cost-utility of health interventions are usually performed through analyses of the incremental impact of the single outcomes on quality-adjusted life years (QALY) and costs associated with the health gain. A threshold of 50000 US$ is considered as cost-effective. A recent review of all published studies that examined cost-effectiveness and cost-utility of Personalized Medicine interventions (1998–2011) showed that the majority (~70%) fall under 50000 US$ per QALY gained, with 20% that are cost saving. This is confirmed by another study that analysed the 47 Personalized-Medicine intervention currently approved (mainly in oncology) in years 2000-2015, which showed, in the majority of the analysed studies, that Personalized Medicine therapies represent a cost-effective/cost-saving treatment option.

Personalised medicine may decrease the average research and development costs for new medicines. Clinical trials are the most expensive part of R&D (nearly 50% of the investment; risen by one third between 2005 and 2007). Biomarkers may enhance the efficacy of clinical trials of new drugs by investing more heavily in early research to identify key biomarkers, and in targeting relevant sub-groups of patients. Smaller (and maybe even shorter) clinical trials are likely to reduce development costs.

CENTER C3 and C7: IMPACT ON THE MARKET AND PUBLIC EXPENDITURE

There are several market segments that may be targeted by the development of rapid tests for genetic traceability of food and food safety. These include, for instance, producers of food ingredients, producers of finished processed food, distributors and suppliers, as well as the mass market retailers (MMR). On the other side, given the important weight on the public health expenditure of foodborne illnesses (detailed below), the development of rapid tests for food safety is expected to produce a large indirect impact on the reduction of public sanitary costs. The innovative tests developed by this HT platform will foster the creation of several new start-ups, specialized in the production and on-demand customization of these technologies to meet the needs of the final users.

Genetic traceability of food

According to Coldiretti, the Italian food industry has a loss of 60 B€/year due to food counterfeiting, which includes both trademark sophistications (false Made in Italy), and substitutions of food ingredients with similar but cheaper food varieties. Additionally, according to the World Customs Organization, food fraud is costing 49 BUS$ annually.
Allied Market Research estimates that the food traceability market is expected to reach 14 BUS$ by 2020, growing with a CAGR of 8.7%. This forecast is based on current technologies that allow for "production chain traceability", such as RFID/RTLS, barcodes, infrared, biometrics and GPS. Genetic traceability of food

F

(varietal identification) represents a different technology that addresses a part of the food traceability market, in particular that regarding food frauds by substitution of an ingredient with a different food variety. Standard technologies that could allow genetic traceability of food are those currently employed in traditional diagnostics, which requires expensive instrumentation or time consuming procedures, and thus are not routinely used for food analysis. Compared to these technologies, the low-cost colorimetric tests for food DNA barcoding, that will be developed in this platform, display several technical advantages (Table 1).

Table 1. *Substitute technologies on the market (adapted from: Wong, E.H.K., et al., Food Research International, 41, 2008, 828.)*

	Applicable to degraded material	Low DNA requirement	Simple protocol	Mixture detection	Time efficient	No prior knowledge required	Reproducible between labs
Hybridization	×			×			
Species-specific primer	×	×	×	×	×		×
RFLP		×	×		×	×	×
SSCP		×			×		
RAPD		×	×		×		
Traditional sequencing	×	×	×		×	×	×
DNA barcoding	×	×	×		×	×	×
Human Technopole	×	×	×	×	×		×

These tests for genetic traceability of food represent, thus, a novel technology with no equivalent on the market, especially in terms of costs. Due to the high market size and absence of substitute products, the tests that will be developed by this HT platform have a huge potential impact on the market, and could create a new, substantial demand for genetic traceability tests, spinning from an existing and unmet market needs.

Food safety

The food safety testing world market is expected to reach 16.1 BUS$ by 2020, according to Markets & Markets (India). Another analysis, from Global Industry Analysts (GIA), predicts an impact of 19.7 BUS$ by 2018.

Only in U.S., according to BCC Research, the food safety testing market will reach 4.3 BUS$ in 2017, increasing at a five-year compound annual growth rate (CAGR) of 5.6%. According to report "Global Markets and Technologies for Food Safety Testing", issued by BCC Research:

"The food safety testing market can be split into five segments based on contaminant type: pathogens, toxins, GMOs (genetically modified organisms), residues, and others. [...]

- *The pathogens segment is expected to increase from nearly US$3 billion in 2012 to US$3.9 billion in 2017, a CAGR of 5.7%.*
- *Toxins are expected to jump from US$141 million in 2012 to $US162 million in 2017, a CAGR of 2.8%.*
- *GMOs, worth US$125 million in 2012, are expected to increase to US$167 million in 2017, a CAGR of 6%.*
- *Residues are expected to climb from US$110 to US$140 million from 2012 to 2017, a CAGR of 4.9%.*
- *The segment made up of other contaminants should increase from US$10 million in 2012 to US$13 million in 2017, a CAGR of 5.4%."*

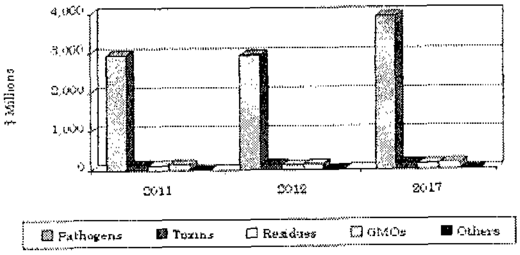


Figure 1. Segmentation and projections for the US food safety testing market (source: BCC Research, Global Markets and Technologies for Food Safety Testing – FOD011G).

Currently, the majority of food safety tests relies on standard instrumental technology, which mostly requires samples shipment to external certified laboratories. This represents an additional and important cost for all the companies involved in the production and distribution of food. Consequently, there is a huge market demand for low cost, rapid tests, that can be performed on field in the food processing/distributing chain.

Moreover, food contamination by foodborne pathogens, occurring in production facilities or along packaging, distribution and storage lines, with the risk of outbreaks of fatal foodborne illnesses, represents a serious public health issue. At least 250 foodborne pathogens have been reported globally (source: ISS, Epicentro). According to WHO (Foodborne Disease Burden Epidemiology Reference Group 2007-2015), approximately 600M of foodborne illnesses have been reported worldwide in 2010, which caused more than 350.000 deaths, and 30.000 cases of various degrees of disability. Moreover, according to "The European Union summary report on trends and sources of *zoonoses, zoonotic* agents and food-borne outbreaks in 2014 (EFSA and ECDC)", 5,251 food-borne outbreaks were reported only in the EU in 2014, which caused more than 6000 hospitalizations and 27 deaths.

The development of low-cost rapid tests for food safety has thus an important impact on the prevention of foodborne epidemics and consequent reduction in connected sanitary expenses. This is exemplified in the following table.

The following table 2, exemplifies the main areas of impact of the HT research in the field, and the main beneficiaries.

Innovation	**Societal gain**								**Beneficiaries**						
	Increased output	Reduced input	Reduced waste	Increased choice	Better quality	Increased competitiveness	Healthier option	Cleaner technology	Agbiotech industry	Food industry	Healthcare industry	Service industry	Farmers	Retailers	Consumers
Bioactives from microrganisms	√	√			√		√		√	√	√		√	√	√
Bioactives from plants	√			√	√	√	√		√	√	√		√	√	√
Bio-control strategies		√					√	√	√	√		√	√	√	√
Cosmeceuticals				√		√	√				√	√		√	√
Crop genomics tools						√			√	√		√			
Decision support systems in agriculture		√	√					√		√		√	√	√	
Fecal transplant				√		√	√				√				√

Fermented products	√			√	√	√	√	√		√	√	√		√	√
Food safety tools					√	√	√			√		√	√	√	√
Functional foods				√	√	√	√			√	√	√		√	√
Health/disease markers				√		√	√				√	√		√	√
Improved crop varieties	√	√	√	√	√	√	√		√			√	√	√	√
Intervention nutrition				√	√	√	√			√	√	√		√	√
Medical foods				√	√	√	√				√	√			√
Metagenomics tools	√	√	√				√	√	√	√	√	√	√		√
Packaging technologies		√	√		√	√	√	√		√				√	√
Personal care				√	√	√	√				√	√		√	√
Personalised nutrition			√	√	√	√	√				√	√	√	√	√
Prebiotics	√			√	√	√	√		√	√	√		√	√	√
Precision agriculture	√	√	√				√	√				√	√		√
Preventive nutrition			√	√	√	√	√			√	√	√		√	√
Probiotic strains				√	√	√	√		√	√	√		√		√
Traceability of food origin					√	√	√			√		√		√	√
Yeast collections	√			√	√	√	√	√	√	√	√	√	√	√	√
food tracability tool					√	√									
Bioprospecting	√	√			√		√								
Lab in a chip for food analysis					√	√									
Food dna barcoding database	√			√	√	√	√	√							
increased food shelf life			√		√		√								

Table 2 Areas of impact of C3 and C7 and main beneficiaries

CENTER C4 and C5 and Facility F3: IMPACT ON THE MARKET AND PUBLIC EXPENDITURE

Centers C4 and C5 will be devoted to the design and the development of novel formalisms, algorithms, and codes, with the goal of obtaining predictive models for life sciences and health. In addition, we envision a well-structured pipeline of activities "from equation to software", and through this pipeline, we seek to develop user-friendly and professional software solutions fostering their widespread distribution and exploitation. These new algorithms can also represent the technological core of high-tech startup companies, enabling them to commercialize software solutions to pharmaceutical, food, and cosmetic industries, in the fields of big data analytics, multiscale computational modeling, and bioinformatics. Small biotech companies will also be reached by these innovative solutions with diverse business models and co-development agreements. In addition, teams of C4 and C5 will apply to H2020 and future framework programs in Health and ICT, as well as to specific technological calls for HPC and data storage infrastructures.

Facility F3 will be devoted to the design and development of a national data repository, which will store administrative, epidemiological, pharmaceutical, clinical, and research data in the domains of health, bioinformatics, and life sciences. This will remarkably improve diagnoses and treatments benefiting the entire health system, including the ultimate goal of precision medicine and personalized treatments for cancer, neurodegenerative diseases, rare diseases, etc.

F

Three KPI could be assumed as indicators of potential economical impact derived by the construction of a national data repository and by the development and application of predictive computational models in life sciences and health: i) optimization of the management of the public health system through a massive informatization of all processes (expected impact of 2-3% on the health system budget); ii) improvement of the reliability of diagnoses and the quality of treatments, which in turn will reduce hospitalization and will impact on the overall population quality of life; iii) development of new enabling technologies thanks to professional data warehousing and computational predictive models, which can have an impact of about 5% on the pharmaceutical and biotech industry-segment that approximately represents the 8% of the national GDP. Table 3 summarizes the expected direct and indirect revenue sources of centers C4 and C5.

Item	Sources
DIRECT REVENUES	
HORIZON 2020	FET (Health and ICT) Center of Excellence for Computing Applications
ERC	YES (three, one ongoing, one starting at Y2, the other starting at Y3)
THER EUROPEAN PROJECTS	Flagship HB
INDUSTRIAL PROJECTS	IBM
LICENSES & PATENT TRANSFER	NA on short term
CHARITIES	Bank Foundations (CRT, CSP)
SOFTWARE – MODELS – ANALYTICS	YES: Open Access
SCIENCE DISSEMINATION	YES: Free
OUTREACH	YES: Free
INDIRECT REVENUES	
START UPS	AI in healthcare (one, starting Y3)
PERSPECTIVE REDUCTION OF THE HEALTH SYSTEM COSTS	National Health System, WHO
CLINICAL TRIALS	Data storage and computational design
PERSPECTIVE REDUCTION OF SOCIAL COSTS INDUCED BY PREDICTIVE MODELS	Public health-care running costs, social costs of diseases: in particular neuro-degenerative diseases and cancer

Table 3 Analysis of the direct and indirect revenue sources of centers C4 and C5 (NA: Not Applicable)

CENTER C6: IMPACT ON THE MARKET AND PUBLIC EXPENDITURE

The main mission of C6 will be to process and analyze high throughput data gathered or generated by Humane Technopole. In addition C6 will manage, explore, and analyze large-scale and high dimensional data on socio-economic decisions and interactions. CADS will integrate and manage massive databases to design and measure the impact of policy decisions on an unprecedented scale, speed, and resolution. Table 1 and 2 synthesize the impact of research activities performed at CADS on attraction of additional resources and on societal benefits. Table 4 exemplifies the main areas of impact of C6 and the main beneficiaries.

Innovation	**Societal Gain**					**Beneficiaries**			
	Sustainability	Predictive capabilities	Industrial Competitiveness	Healty ageing	Human capital	Public Sector	Healthcare industry	Service industry	Citizens
Data integration and analytics for policies	√	√	√	√	√	√	√	√	√
Analytics for health and the life sciences		√	√	√		√	√	√	√
Foundational data science		√			√	√	√	√	√
Analysis of industrial and financial systems		√	√		√	√	√	√	√
From now casting to long term projections	√	√	√	√	√	√	√	√	√
Management systems and solutions	√		√	√	√	√	√	√	√

Table 4 Areas of impact of C6 and main beneficiaries

Per Copia conforme