Finding Consistent Disease Subnetworks

Limsoon Wong

26 March 2010

(Joint work with Donny Soh, Difeng Dong, Yike Guo)

Plan

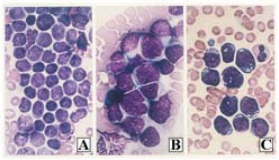
- **An issue in gene expression analysis**
- **Comparing pathway sources**
  - Comprehensiveness
  - Consistency
  - Compatibility
- **Matching pathways in different sources**
- **Finding more consistent disease subnetworks**

An Issue in Gene Expression Analysis

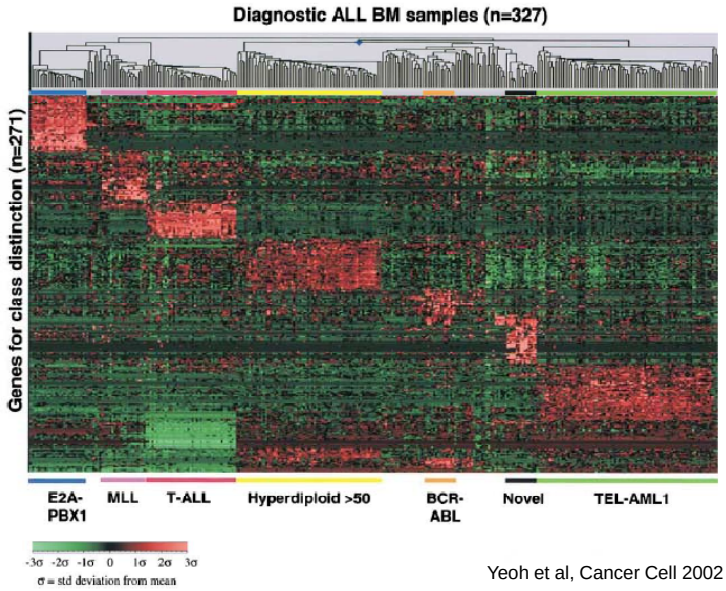
First, the good news..

Childhood Acute Lymphoblastic Leukemia

- **Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid>50**
- **Diff subtypes respond differently to same Tx**
- **Over-intensive Tx**
  - Development of secondary cancers
  - Reduction of IQ
- **Under-intensiveTx**
  - Relapse
- **The subtypes look similar**

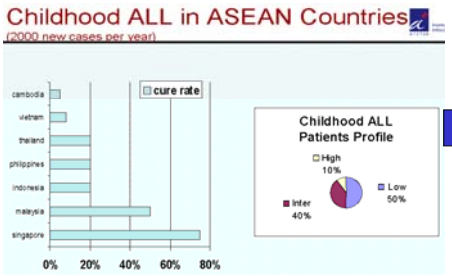


- **Conventional diagnosis**
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics
  - ⇒Unavailable in developing countries



Yeoh et al, Cancer Cell 2002

Impact



Conventional Tx:
- intermediate intensity to all

⇒ 10% suffers relapse
⇒ 50% suffers side effects
⇒ costs US$150m/yr

Our optimized Tx:
- high intensity to 10%
- intermediate intensity to 40%
- low intensity to 50%
- costs US$100m/yr

High cure rate of 80%
• Less relapse
• Less side effects
• Save US$51.6m/yr

Now, the bad news..

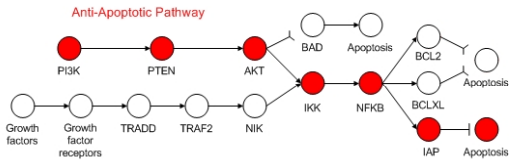
Percentage of Overlapping Genes

- **Low % of overlapping genes from diff expt in general**
  - Prostate cancer
    - **Lapointe et al, 2004**
    - **Singh et al, 2002**
  - Lung cancer
    - **Garber et al, 2001**
    - **Bhattacharjee et al, 2001**
  - DMD
    - **Haslett et al, 2002**
    - **Pescatori et al, 2007**

Datasets	DEG	POG
Prostate Cancer		
	Top 10	0.30
	Top 50	0.14
	Top100	0.15
Lung Cancer		
	Top 10	0.00
	Top 50	0.20
	Top100	0.31
DMD		
	Top 10	0.20
	Top 50	0.42
	Top100	0.54

Zhang et al, Bioinformatics, 2009

Gene Regulatory Circuits



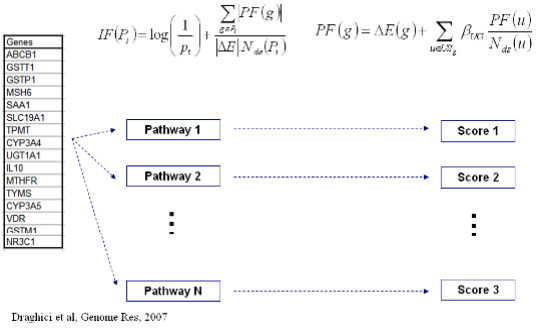
- **Each disease subtype has underlying cause**
- **There is a unifying biological theme for genes that are truly associated with a disease subtype**

- **Uncertainty in selected genes can be reduced by considering biological processes of the genes**
- **The unifying biological theme is basis for inferring the underlying cause of disease subtype**

Towards More Meaningful Genes

- **ORA**
  - Khatri et al
  - *Genomics*, 2002
- **FCS**
  - Pavlidis & Noble
  - *PSB* 2002
- **GSEA**
  - Subramanian et al
  - *PNAS*, 2005
- **Pathway Express**
  - Draghici et al
  - *Genome Res*, 2007

Gene Class Testing: Pathway Express



Draghici et al, Genome Res, 2007

All of these newer methods rely on gene group or pathway information.

But how good are the available sources of pathway information?

Comparing Pathway Sources

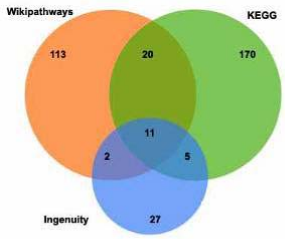
Data Sources

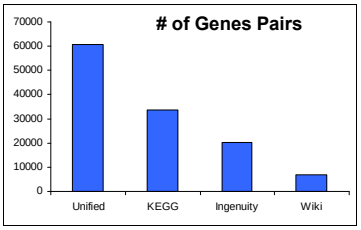
- **KEGG**
  - Curated by a single lab
  - Long famous history
  - Used by many people
- **Wikipathways**
  - Community effort
  - new curation model
- **Ingenuity**
  - Commercial effort
  - Used by many biopharma's

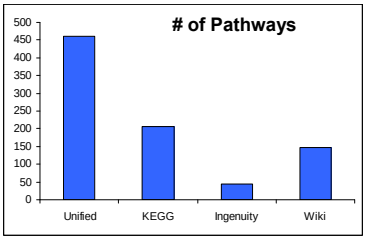
INGENUITY® SYSTEMS

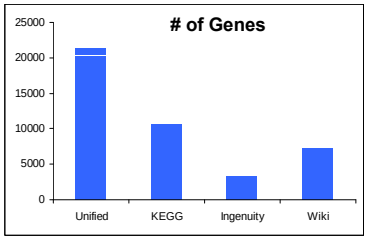
15

Low Comprehensiveness of Pathway Sources





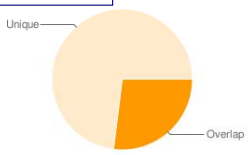




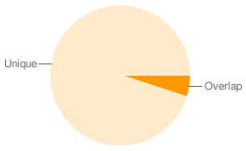
16

Low Consistency of Pathway Sources

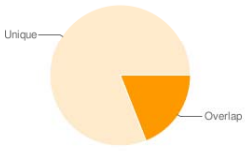
Gene Pair Overlap



Wiki vs KEGG

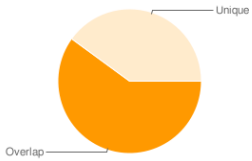


Wiki vs Ingenuity

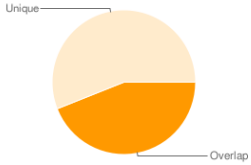


KEGG vs Ingenuity

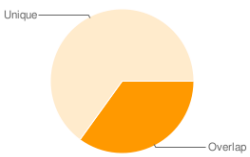
Gene Overlap



Wiki vs KEGG



Wiki vs Ingenuity



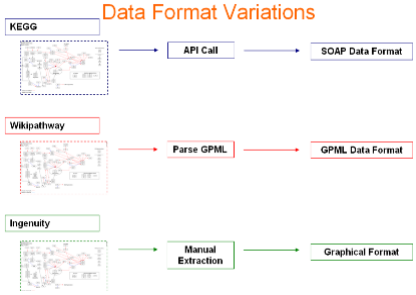
KEGG vs Ingenuity

Example: Apoptosis Pathway

Apoptosis Pathway			
	Wiki x KEGG	Wiki x Ingenuity	KEGG x Ingenuity
Gene Pair Count:	144 vs 172	144 vs 3557	172 vs 3557
Gene Count:	85 vs 80	85 vs 176	80 vs 176
Gene Overlap:	38	28	30
Gene % Overlap:	48%	33%	38%
Gene Pair Overlap:	23	14	24
Gene Pair % Overlap:	16%	10%	14%

Would Unifying Pathway Sources Help?

- **Incompatibility Issues!**



- **Data extraction method variations**
- **Format variations**
- **Data differences**
- **Pathway name differences**
- **Gene/GeneID name differences**

The preceding analyses hide an intricate issue…

The same pathways in the different sources are often given different names.

So how do we even know two pathways are the same and should be compared / merged?

Intricacy of Pathway Matching

Possible Ways to Match Pathways

- **Match based on name**
  - Pathways w/ similar name should be the same pathway
  - But annotations are very noisy
  - ⇒Likely to mismatch pathways?
  - ⇒Likely to match too many pathways?
- **Are the followings good alternative approaches?**
  - Match based on overlap of genes
  - Match based on overlap of gene pairs

Matching Pathways by Name

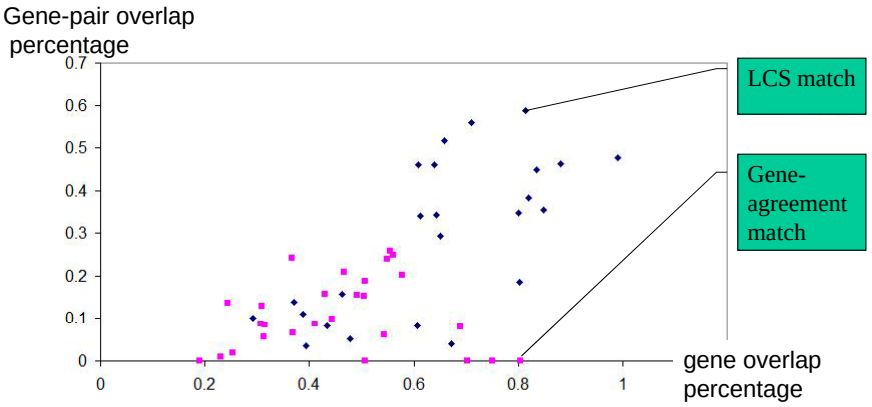
- **LCS procedure**
  - Given pathway X in db A
  - Sort pathways in db B by "longest common substring" with X
  - Manually scan the ranked list to choose closest nomen-clatural match
- **Issue: Accuracy**
  - When LCS says two pathways are the same one, are they really the same?
- **Issue: Completeness**
  - When LCS says two pathways are different, are they really different?

LCS vs Gene-Agreement Matching

- **Accuracy**
  - 94% of LCS matches are in top 3 gene agreement matches
  - 6% of LCS matches not in top 3 of gene agreement matches; but their gene-pair agreement levels are higher
- **Completeness**
  - Let Pi be pathway in db A LCS cannot find match in db B
  - Let Qi be pathway in db B with highest gene agreement to Pi
  - Gene-pair agreement of Pi-Qi is much lower than pathway pairs matched by LCS

LCS is better than gene-agreement based matching!

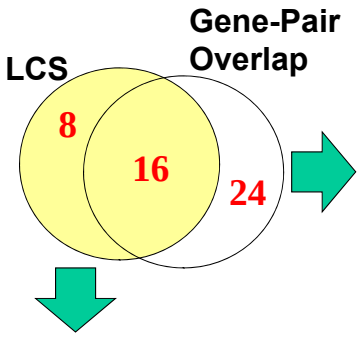
LCS vs Gene-Agreement Matching



- **LCS consistently has higher gene-pair agreement**

⇒ **LCS is better than gene-agreement based matching!**

LCS vs Gene-Pair Agreement Matching



ErbB signaling pathway	JAK/Stat Signaling
Calcium signaling pathway	Synaptic Long Term Potentiation
Apoptosis	Toll-like receptor signaling pathway
VEGF signaling pathway	Axonal Guidance Signaling
Gap junction	PPAR-alpha/RXR-alpha Signaling
Natural killer cell mediated cytotoxicity	Fc Epsilon RI Signaling
T cell receptor signaling pathway	Axonal Guidance Signaling
B cell receptor signaling pathway	Axonal Guidance Signaling
Olfactory transduction	cAMP-mediated Signaling
GnRH signaling pathway	B Cell Receptor Signaling
Melanogenesis	Wnt Signaling Pathway and Pluripotency
Type II diabetes mellitus	Insulin Recpetor Signaling
Colorectal cancer	Toll-like receptor signaling pathway
Renal cell carcinoma	Axonal Guidance Signaling
Pancreatic cancer	PTEN Signaling
Endometrial cancer	PTEN Signaling
Glioma	ERK/MAPK Signaling
Prostate cancer	JAK/Stat Signaling
Basal cell carcinoma	Wnt Signaling Pathway and Pluripotency
Melanoma	FGF Signaling
Chronic myeloid leukemia	GM-CSF Signaling
Acute myeloid leukemia	PTEN Signaling
Small cell lung cancer	Toll-like receptor signaling pathway
Non-small cell lung cancer	GM-CSF Signaling

The 24 pathway pairs singled out by maximal gene-pair overlap

Regulation of actin cytoskeleton	Regulation of Actin Cytoskeleton
Wnt signaling pathway	Wnt Signaling Pathway
T cell receptor signaling	t cell receptor Signaling
VEGF signaling	VEGF Signaling
MAPK signaling	MAPK Cascade
Apoptosis	Apoptosis
Apoptosis	Apoptosis Signaling
Toll-like receptor	Toll-like receptor signaling pathway

The 8 pathway pairs singled out by LCS

Note: We consider only pathway pairs that have at least 20 reaction overlap.

LCS vs Gene-Pair Agreement Matching

- **Gene-pair agreement match will miss when**
  - Pathway P in db A has few overlap with pathway P in db B due to incompleteness of db, even if pathway name matches perfectly!
  - Example: wnt signaling pathway, VEGF signaling pathway, MAPK signaling pathway, etc. in KEGG don't have largest gene-pair overlap w/ corresponding pathways in Wikipathways & Ingenuity

⇒ **Bad for getting a more complete unified pathway P**

LCS vs Gene-Pair Agreement Matching

- **Pathways having large gene-pair overlap are not necessarily the same pathways**
- **Examples**
  - "Synaptic Long Term Potentiation" in Ingenuity vs "calcium signalling" in KEGG
  - "PPAR-alpha/RXR-alpha Signaling" in Ingenuity vs "TGF-beta signaling pathway" in KEGG

⇒ **Difficult to set correct gene-pair overlap threshold to balance against false positive matches**

Bad Gene-Pair Agreement Matches

- **"Synaptic Long Term Potentiation" in Ingenuity vs "calcium signalling" in KEGG**
  - **Calcium signaling pathway in KEGG describes general mechanism of external calcium signal transduction into cells**
  - **Calcium signal transduction can activate multiple downstream pathways, LTP is one of them**
  - ⇒ **LTP in Ingenuity is only a downstream event of the calcium pathway in KEGG**
- **"PPAR-α/RXR- α signaling" in Ingenuity vs "TGF-β signaling" in KEGG**
  - **PPARα/RXRα plays essential roles in the regulation of cellular differentiation, development, metabolism, and tumorigenesis**
  - **TGF-β acts as antiproliferative factor in normal cells at early stages of oncogenesis**
  - ⇒ **They are independent. The reason they are paired is that they have a mutual inhibition**

- **Having found a good way to match up pathways in different datasources, we proceeded to build a big unified pathway db….**

PathwayAPI
= KEGG
+ Wikipathways
+ Ingenuity

More Consistent Disease Subnetworks

But these methods still don't return the precise parts of a pathway that are significant…

- **ORA**
  - Khatri et al
  - Genomics, 2002
- **FCS**
  - Pavlidis & Noble
  - PSB 2002
- **GSEA**
  - Subramanian et al
  - PNAS, 2005
- **Pathway Express**
  - Draghici et al
  - Genome Res, 2007

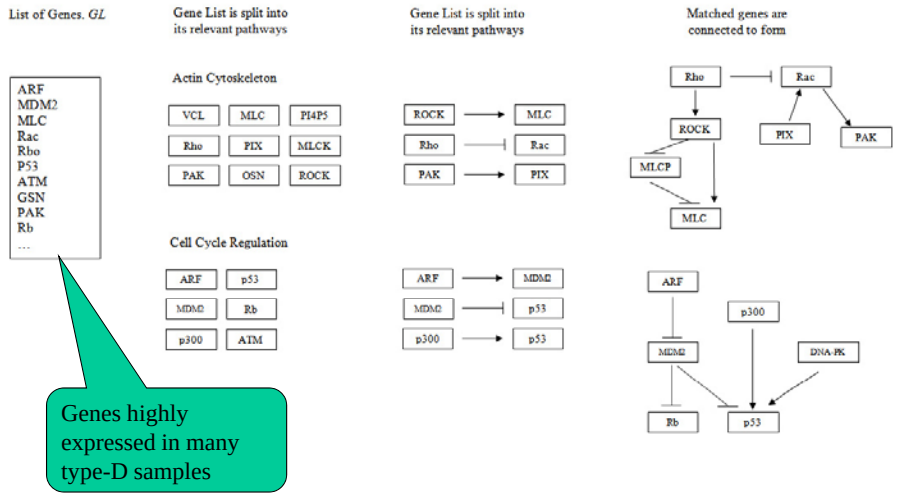
Test whole gene group at a time

Test a node and its immediate neighbours

The SNet Method

- **Group samples into type D and ¬D**
- **Extract & score subnetworks for type D**
  - Get list of genes highly expressed in most D samples
    - **These genes need not be differentially expressed!**
  - Seggregate these genes into pathways
  - Locate largest connected components (ie., candidate subnetworks) from these pathway graphs
  - Score each subnetwork
- **Repeat the same on ¬D samples**
- **T-test on the two sets of scores to get significant subnetworks for D**

SNet: Extract Subnetworks



SNet: Score Subnetworks

Step 2: Subnetwork Scoring We assign a score vector $SN^{v\_score}_{sn,d}$ with respect to phenotype d to each subnetwork sn within SN^{List} according to Equation 1.

$$SN^{v\_score}_{sn,d} = \langle SN^{i\_score}_{sn,1,d}, SN^{i\_score}_{sn,2,d}, ..., SN^{i\_score}_{sn,n,d} \rangle \quad (1)$$

Where n is the number of patients in phenotype d. The formula $SN^{i\_score}_{sn,i,d}$ for the i^{th} patient (also the i^{th} element of this vector) is given by:

$$SN^{i\_score}_{sn,i,d} = \sum_{j=1}^{g} G^{score}_{sn,j,d} \quad (2)$$

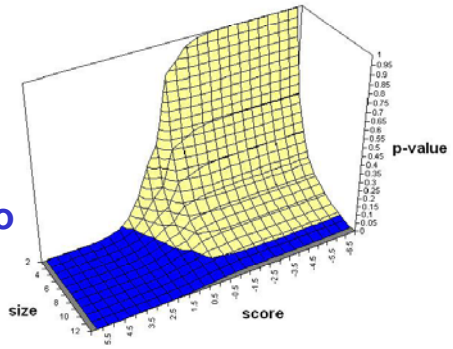
$G^{score}_{sn,j,d}$ refers to the score of the j^{th} gene (say, gene x) in the subnetwork sn for phenotype d. (This score $G^{score}_{sn,j,d}$ is given by Equation 3) and is simply given by:

$$G^{score}_{sn,j,d} = k/n \quad (3)$$

Where k is the number of patients of phenotype d who has gene x highly expressed (top $\alpha\%$) and n is the total number of patients of phenotype d. The entire Step 2 is repeated for the other disease phenotype $\neg d$, giving us the score vectors, $SN^{v\_score}_{sn,d}$ and $SN^{v\_score}_{sn,\neg d}$ for the same set of connected components. The t-test is finally calculated between these two vectors, creating a final t-score for each subnetwork sn within SN_{List}.

SNet: Significant Subnetworks

- **Randomize patient samples many times**
- **Get t-score for subnetworks from the randomizations**
- **Use these t-scores to establish null distribution**
- **Filter for significant subnetworks from real samples**



Let's see whether SNet gives us subnetworks that are

(i) more consistent between datasets of the same types of disease samples

(ii) larger and more meaningful

Recall Examples from "Bad News"

- **Low % of overlapping genes from diff expt in general**
  - Prostate cancer
    - **Lapointe et al, 2004**
    - **Singh et al, 2002**
  - Lung cancer
    - **Garber et al, 2001**
    - **Bhattacharjee et al, 2001**
  - DMD
    - **Haslett et al, 2002**
    - **Pescatori et al, 2007**

Datasets	DEG	POG
Prostate Cancer		
	Top 10	0.30
	Top 50	0.14
	Top100	0.15
Lung Cancer		
	Top 10	0.00
	Top 50	0.20
	Top100	0.31
DMD		
	Top 10	0.20
	Top 50	0.42
	Top100	0.54

Zhang et al, Bioinformatics, 2009

Better Subnetwork Overlap

Table 1. Table showing the percentage overlap significant subnetworks between the datasets. Each row refers to a separate disease (as indicated in the first column). Each disease is tested against two datasets depicted in the second and third column. The overlap percentages refer to the pathway overlaps obtained from running SNet (column 4) and GSEA (column 5) The actual number of overlaps are parenthesized in the same columns.

Disease	Dataset 1	Dataset 2	SNet	GSEA
Leuk	Golub	Armstrong	83.3% (20)	0.0% (0)
Subtype	Ross	Yeoh	47.6% (10)	23.1% (6)
DMD	Haslett	Pescatori	58.3% (7)	55.6% (10)
Lung	Bhatt	Garber	90.9% (9)	0.0% (0)

- **For each disease, take significant subnetworks from one dataset and see if it is also significant in the other dataset**

Better Gene Overlaps

Table 2. Table showing the number and percentage of significant overlapping genes. γ refers to the number of genes compared against and is the number of unique genes within all the significant subnetworks of the disease datasets. The percentages refer to the percentage gene overlap for the corresponding algorithms.

Disease	γ	SNet	GSEA	SAM	t-test
Leuk	84	91.3%	2.4%	22.6%	14.3%
Subtype	75	93.0%	4.0%	49.3%	57.3%
DMD	45	69.2%	28.9%	42.2%	20.0%
Lung	65	51.2%	4.0%	24.6%	26.2%

- **For each disease, take significant subnetworks extracted independently from both datasets and see how much their genes overlap**

Larger Subnetworks

Table 3. Table comparing the size of the subnetworks obtained from the t-test and from SNet. The first column shows the disease and the second column shows the number of genes which comprised of the subnetworks. The third and fourth column depicts the number of genes present within each subnetwork for the t-test and SNet respectively. So for instance in the leukemia dataset, we have 8 subnetworks with size 2 genes, 1 subnetwork with size 3 genes for the t-test. For SNet, we have 2 subnetworks with size 5 genes, 3 subnetworks with size 6 genes, 2 subnetworks with size 7 genes and 1 subnetwork with a size of ≥ 8 genes

Disease	γ	Num Genes (t-test)				Num Genes (SNet)			
		2	3	4	5	5	6	7	≥ 8
Leuk	84	8	1	0	0	2	3	2	1
Subtype	75	5	1	1	1	1	0	1	6
DMD	45	3	1	0	0	1	0	0	5
Lung	65	3	2	1	0	5	3	0	1

Remarks

What have we learned?

- **Significant lack of concordance betw db's**
  - Level of consistency for genes is 0% to 88%
  - Level of consistency for genes pairs is 0%-61%
  - Most db contains less than half of the pathways in other db's
- **Matching pathways by name is better than matching by gene overlap or gene-pair overlap**
- **SNet method yields more consistent and larger disease subnetworks**

Acknowledgements

Donny Soh

Difeng Dong

Yike Guo

- A*STAR AIP scholarship
- A*STAR SERC PSF grant

Agency for Science, Technology and Research
SINGAPORE