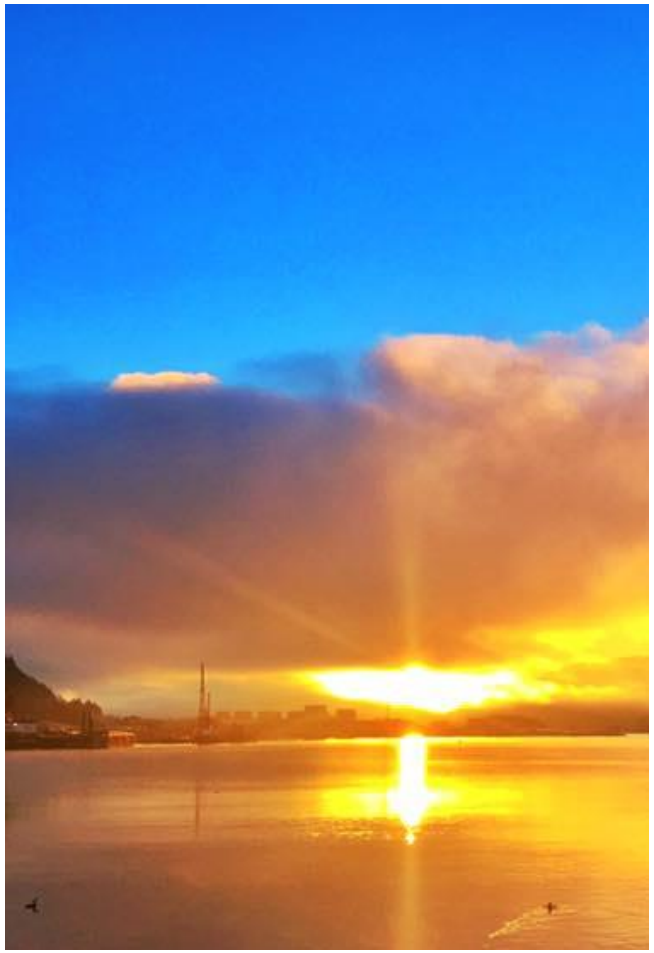
Experimenting with experiments: what can we learn from *small N, big p* biomarker studies?

Fabio Rigat (GSK, Clinical Biostatistics)



Disclaimer and acknowledgement

This work reflects the author's views, not GSK's position

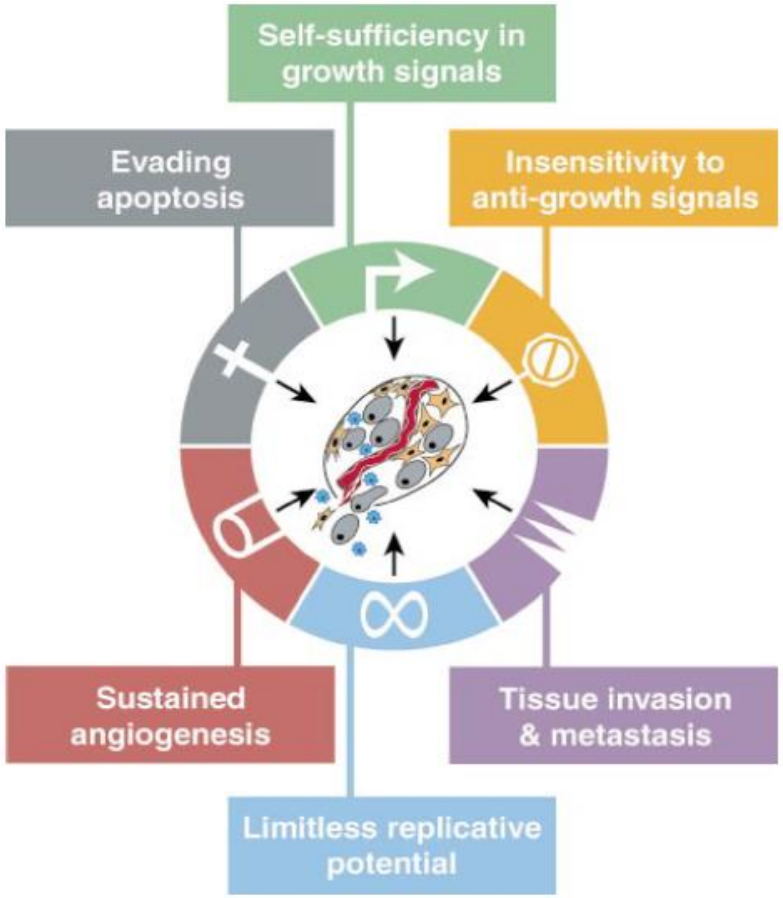
We are grateful to all patients who consent to the bone marrow biopsies enabling the generation of clinical gene expression data

Outline

- **Hallmarks of cancer**: alterations in cell physiology characterising tumorigenesis
  - Tumour baseline gene expression data analysis: annotation-naïve and annotation-based
    - Gene annotation primer for statisticians
- **Association of baseline RNA levels to clinical responses**
  - Annotation-based regression model
  - OCs of annotation-naïve and annotation-based inferences using synthetic data
- **Summary and way forward**
  - Annotation-based regression enables
    - interpretation of treatment effects by known pathways and biological processes
    - planning of joint analysis of the multiple studies needed to accrue larger data bases

Dysregulated gene expression can lead self-sufficient cell growth

- Hanahan et al, Cell, 2000 (in figure and B/U): "*We suggest that […] cancer […] is a manifestation of six essential alterations in cell physiology*"
  - Epigenetic dysregulation is one of the mechanisms mediating alterations in cell physiology of cancer cells**, as** dysregulated DNA transcription into RNA leads to self-sufficient growth
- **Case study:**
  - **RNA** extracted from bone marrow of *N = few* PhI trial subjects at baseline
  - **RNAseq** quantifies expression of *p = tens of thousands* of transcripts by shotgun sequencing, counting reads mapped to the transcriptome assembly
    - **can *p>>N* designs provide any evidence of predictive associations?**
- **Note:** *establishing predictive biomarkers has been historically challenging*, due to few good hypotheses, weak trial designs and simplistic data analysis methods



The six dimensions of tumorigenesis in Hanahan et al, Cell, 2000. Currently 10 dimensions have been identified

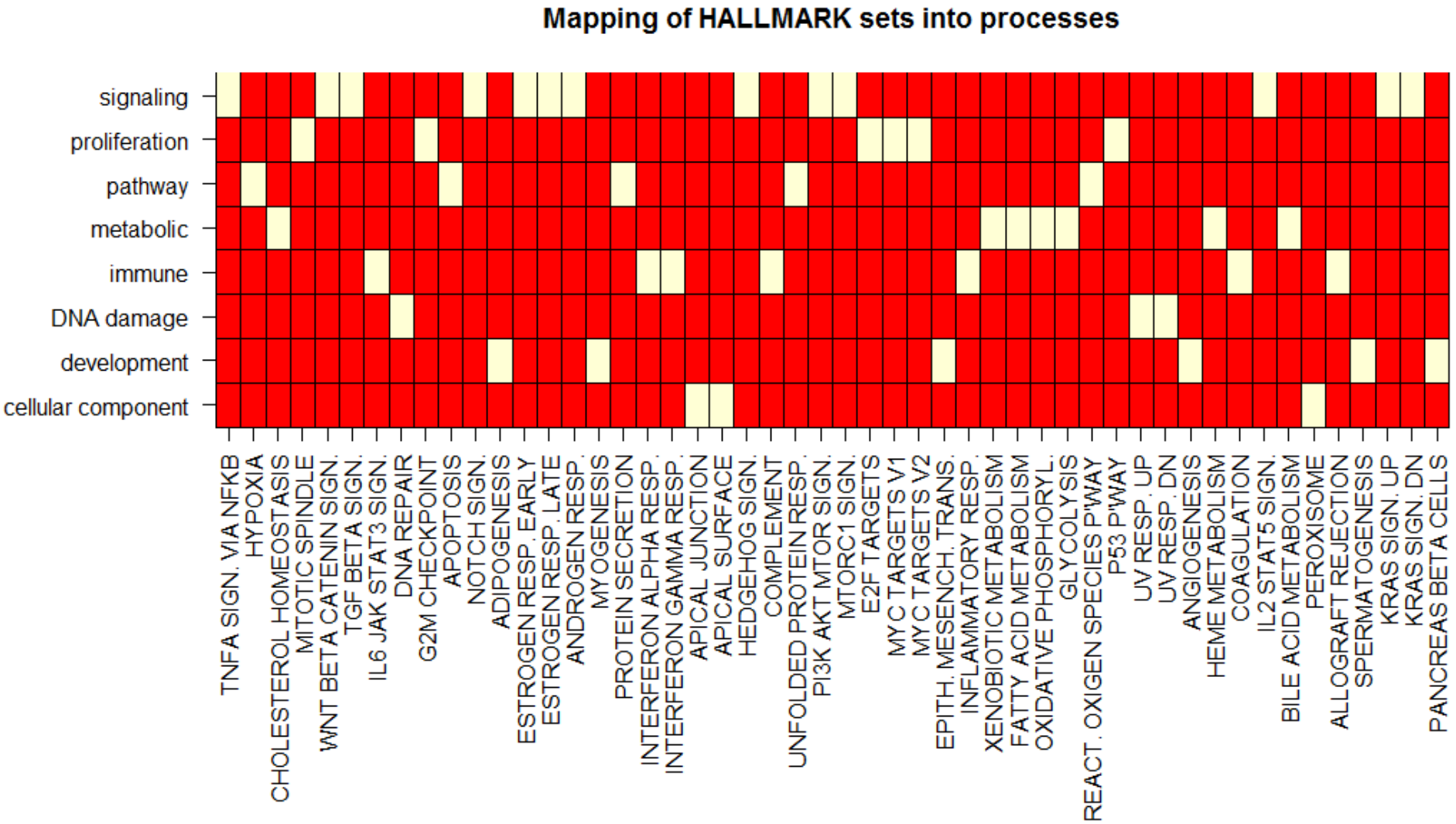
Annotation-naïve and annotation-based analysis of patient-level RNAseq data

- **RNAseq data analysis challenges**
  1. ***N* << *p***: "a stack of 3644 double-deckers"
  2. **Measurability**: few transcripts > LLOQ in all subjects
  3. **Sample composition**: the % of cancer cells in RNA samples likely to show presence of non-tumour cells

- **Two avenues for learning from RNAseq data:**
  - **Bottom-up**
    - For each RNA transcript, *test* responders vs non-responders
    - Map the test statistics into pathways for interpretation (GEA)
  - **Top-down**
    - Map the annotated RNA transcripts into pathways
    - *Estimate* the probability that different pathway expressions result in different clinical outcomes via regression
      - Regression allows accounting for potential predictions other than RNAseq, e.g. **dose, trial arm etc**.

	Analysis options	
	Bottom-up	**Top-down**
+	Off-the-shelf	Phrases quantitative decision rules using pathways
-	No built-in interpretive framework	Few transcripts annotated Needs tailor-made model

Gene annotation databases are a working framework for interpreting empirical associations with expression

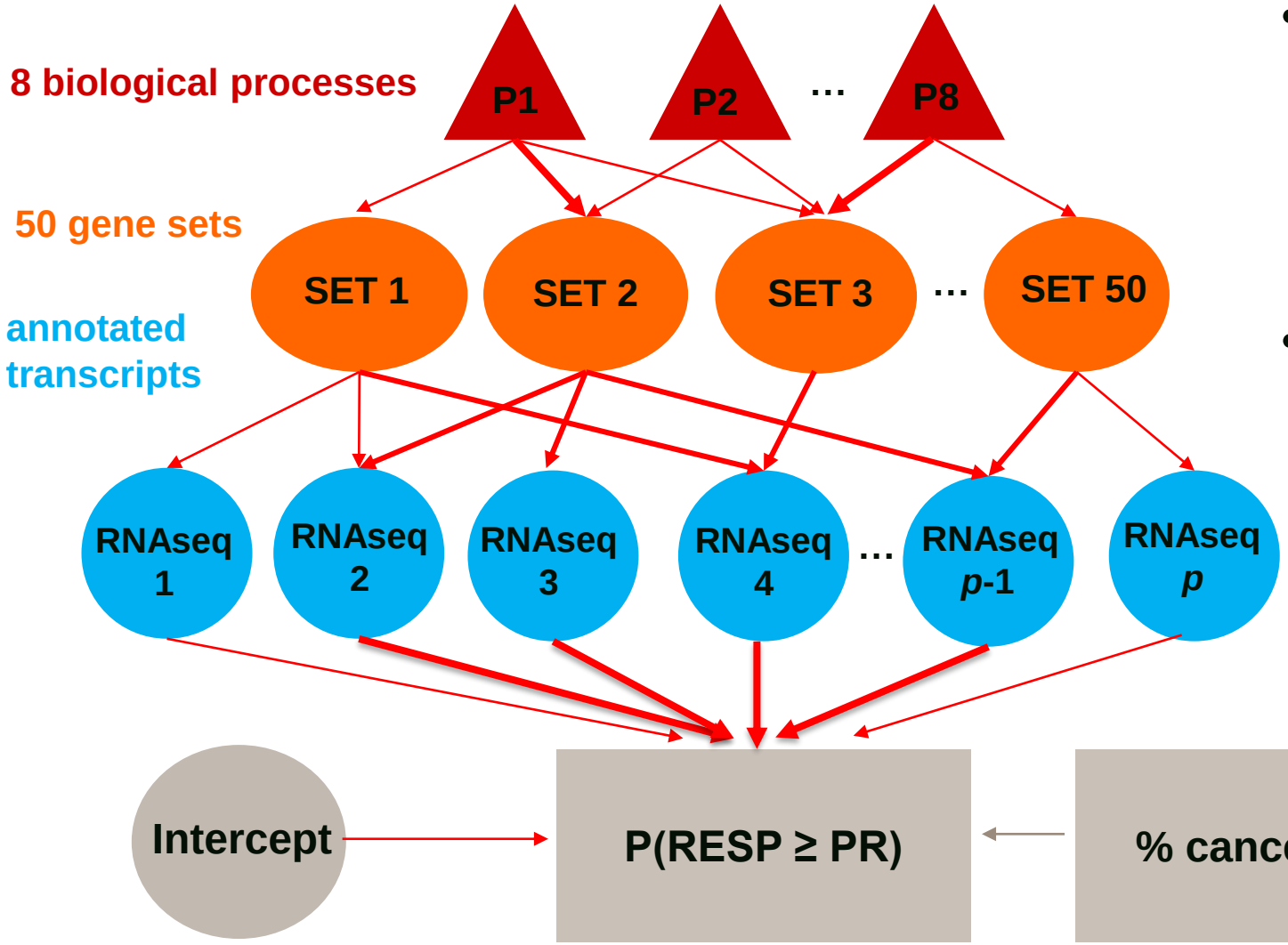


- The molecular signatures database collections (**MSigDB**) is a simple framework for illustrating the role of functional annotations in clinical biomarkers data analysis
- **Relatively few measurable RNAseq transcripts** are annotated in MSigDB

http://geneontology.org/page/about

http://software.broadinstitute.org/gsea/msigdb/collections.jsp

Gene annotations wire a systems biology model for interpreting associations of expression and clinical responses



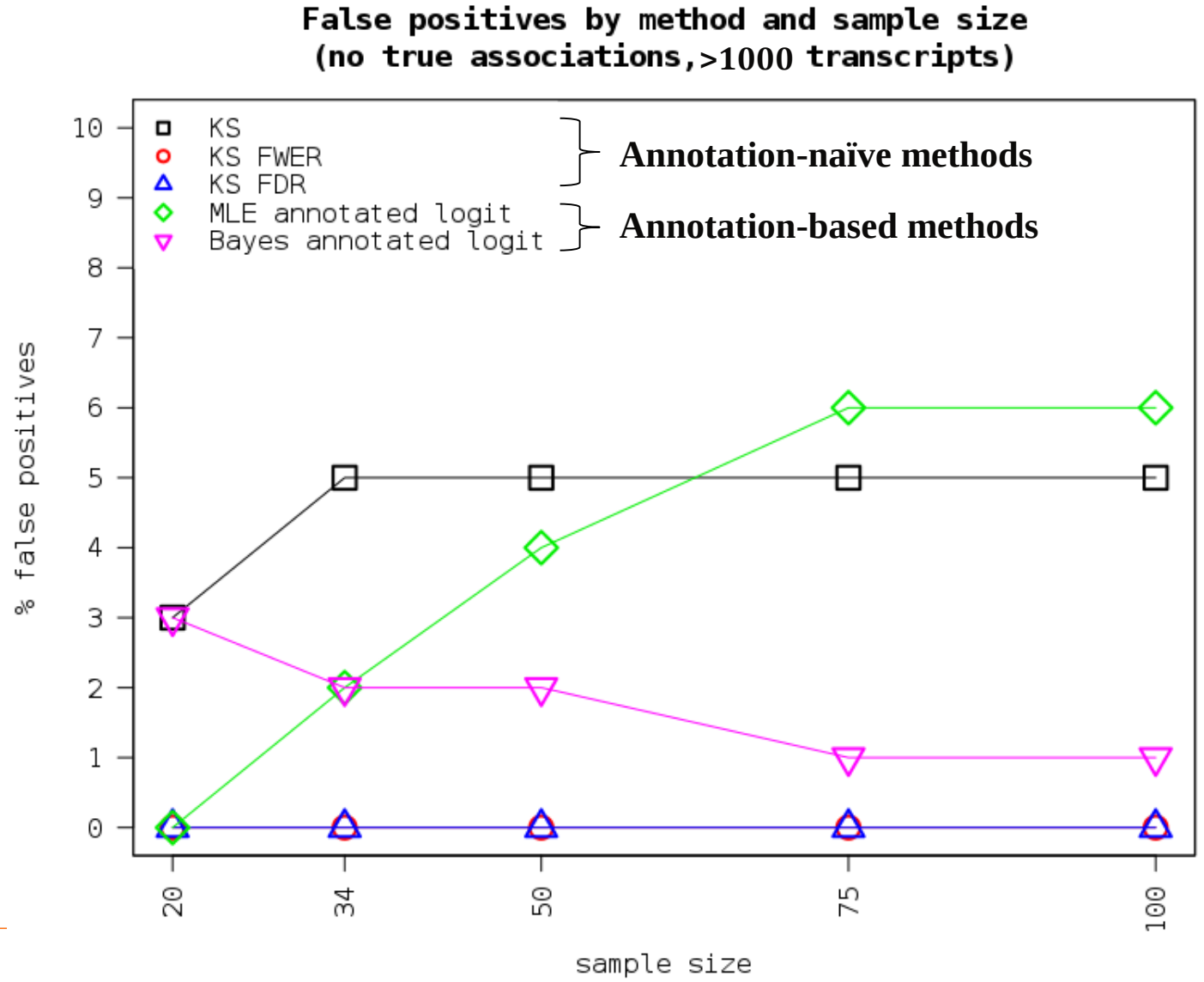
- **What is known:**
  - system wiring (annotations)
  - RNAseq FPKM (potential predictors)
  - Sample quality (% cancer cells)
  - Clinical response (RESP)
- **What is estimable from the data:**
  - **strength of associations**

	Estimation options	
	MLE	**Bayes**
+	Fast numerical estimation	Estimates process, set and transcripts coefficients
-	Estimates process coefficients	Prior calibration and sensitivity

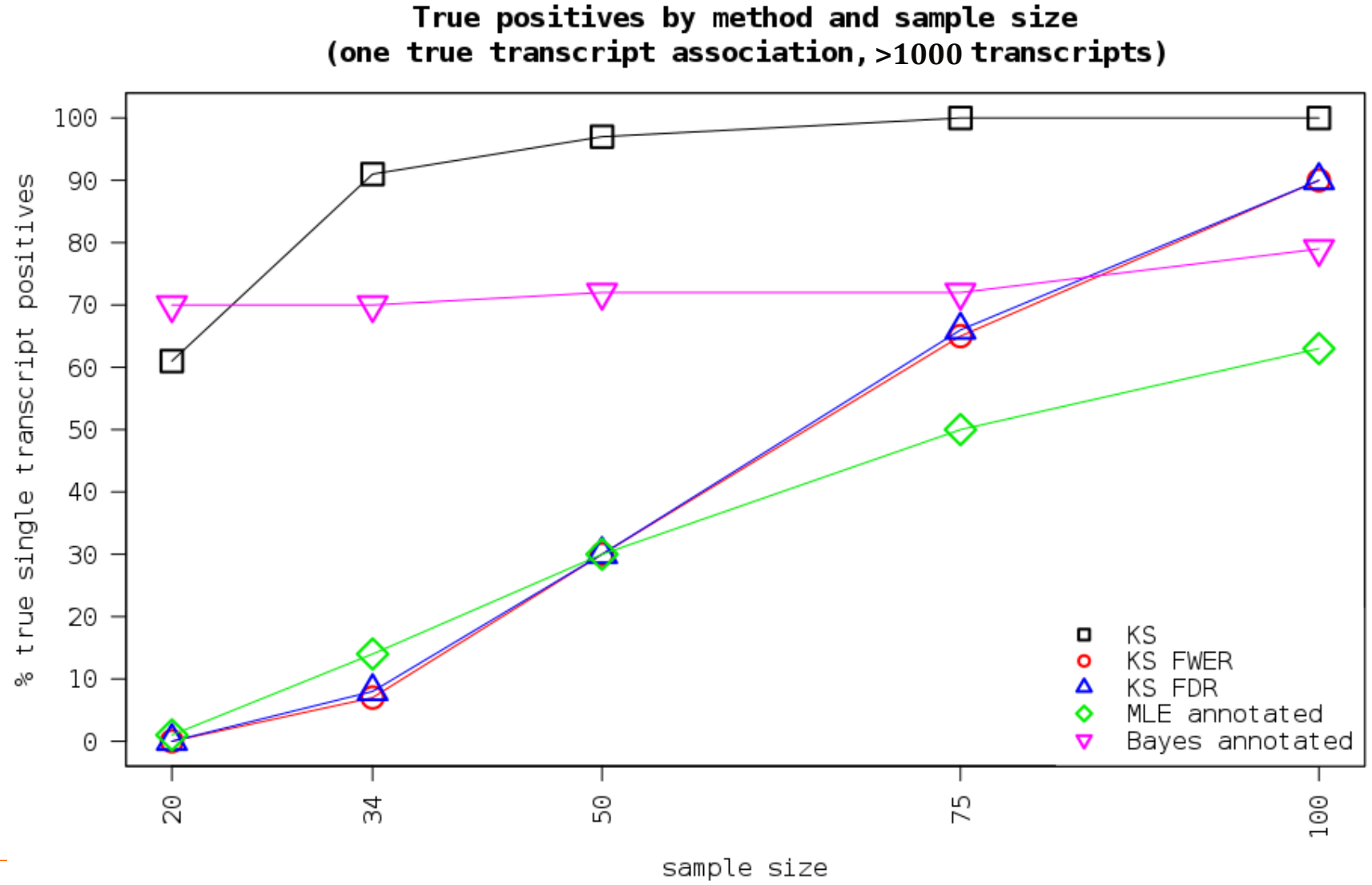
Performance of data analysis models is compared under a *set truth* prior to real data analysis

- **Data simulation** : $Y_i \sim^{i.i.d.} Bernoulli(\pi)$, $X_{ij} \sim^{ind} LogN(\mu_j, \sigma_0 = 1)$ for i $= 1, \ldots, N$ *and* $j = 1, \ldots, p$
  Null hypothesis: $\mu_j = \mu_0 = 6.2$
  One true transcript association: $\mu_j := \mu_0 \times 1_{\{j \neq j_{true}\}} + 4.9 \times 1_{\{Y_i=0, j=j_{true}\}} + 7.7 \times 1_{\{Y_i=1, j=j_{true}\}}$
  One true process association: $\mu_j := \mu_0 \times 1_{\{j \; not \; in \; P_{true}\}} + 4.9 \times 1_{\{Y_i=0, j \in P_{true}\}} + 7.7 \times 1_{\{Y_i=1, j \in P_{true}\}}$
- **For each simulated dataset *s* of sample size *N* calculate**:
  1. $p_{sj}(N)$: the KS p-value between responders and non-responders and $d_{sj}(N) = 1_{\{p_{sj}(N)<0.05\}}$
  2. $p_{sj}^{Holm}(N)$: Holm's FWER p-value p_j and $d_{sj}^{Holm}(N) = 1_{\{p_{sj}^{Holm}(N)<0.05\}}$
  3. $p_{sj}^{PH}(N)$: Benyamini-Hockberg's FDR p_j-value and $d_{sj}^{BH}(N) = 1_{\{p_{sj}^{BH}(N)<0.05\}}$
  4. $(L_{sk}(N), U_{sk}(N))$ the 95%CRI of process $k = 1, \ldots, 8$ and $d_{sk}^{BAR}(N) = 1_{\{0 \in (L_{sk}(N), U_{sk}(N))\}}$
- **Calculate operating characteristics** at N by averaging $d_{sj}(N), d_{sj}^{Holm}(N), d_{sj}^{BH}(N), d_{sk}^{BAR}(N)$ over *s*, i.e.
  False positives: proportion of transcripts or processes falsely associated to response
  True positives: proportion of transcripts or processes truly associated to response
- **"Don't do this in excel"**: simulations run in parallel on >150 CPUs for tolerable run times

All methods control false positive RNAseq association probabilities



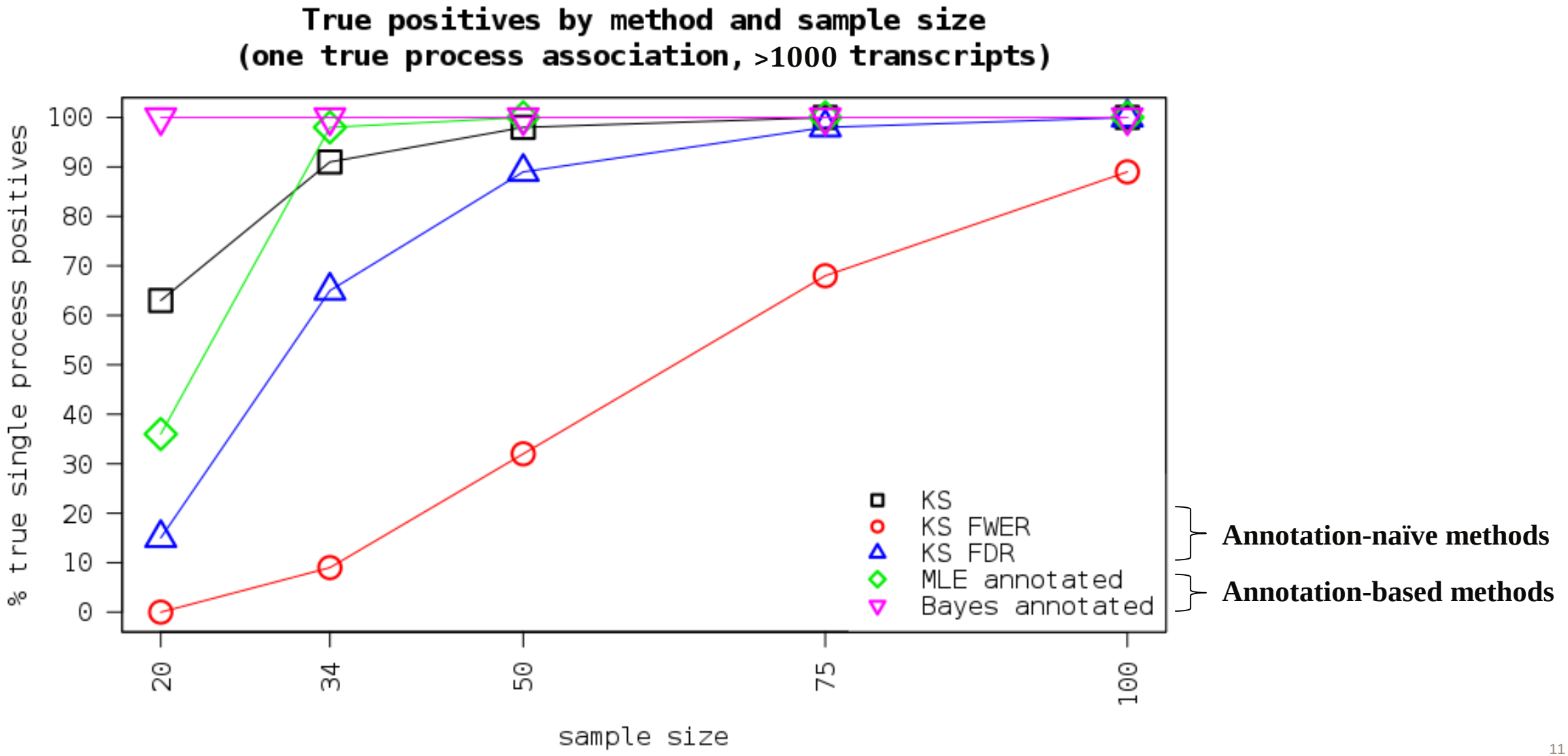
Annotation-naïve methods have greater power for detecting a single transcript RNAseq association



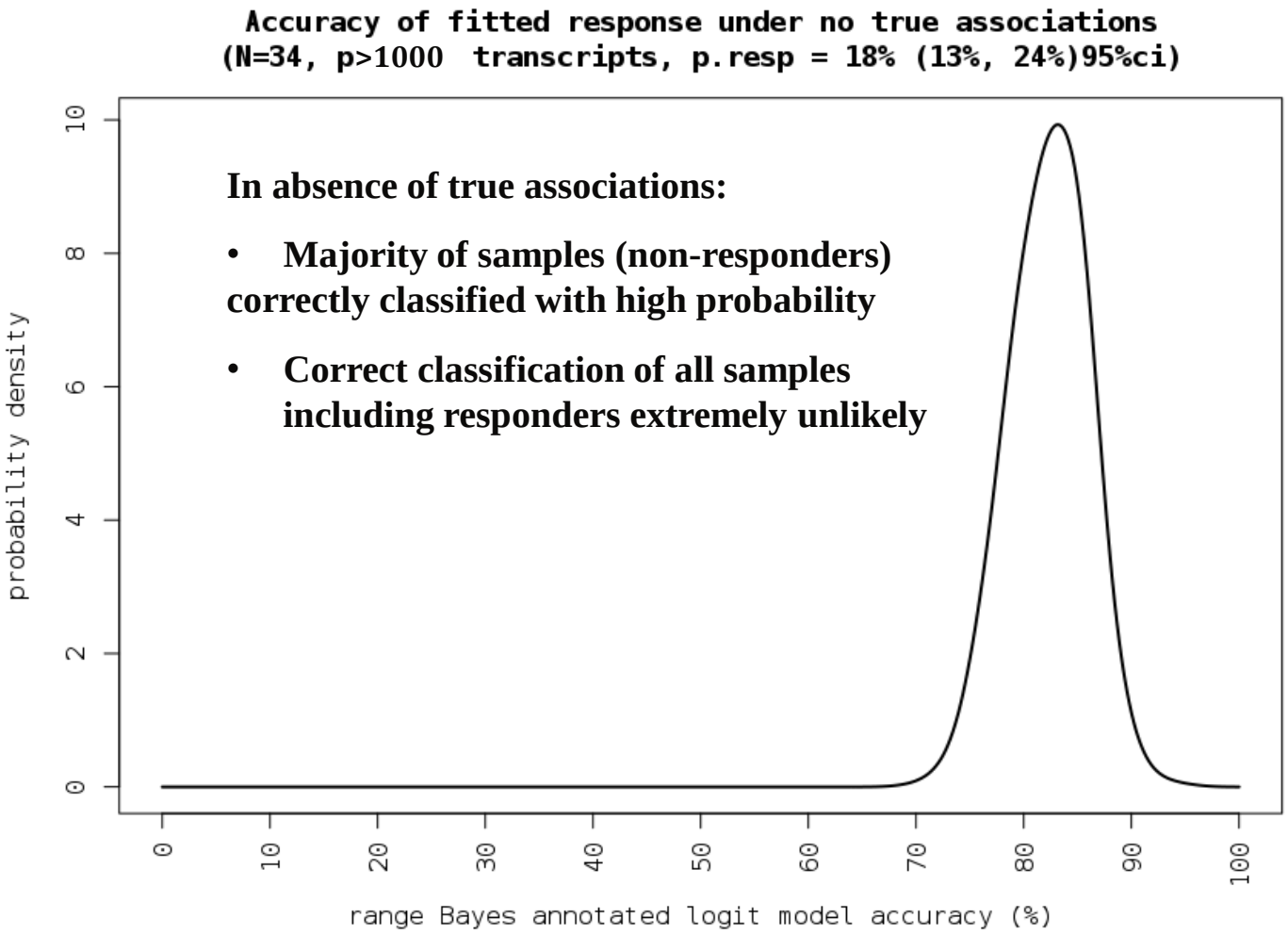
Annotation-naïve methods

Annotation-based methods

Annotation-based regression has greater power for detecting a single process RNAseq association



Annotation-based analysis of case study data



- **Distribution of classification accuracy of annotation-based model under no association** (in figure) shows that observed accuracy of fit (100%) is statistically significant
  - Compared to no association, the baseline RNAseq data enable the regression model to correctly identify all responders
- **95% CRIs of all pathways and processes regression coefficient estimates do not show strong evidence** of associations
  - Strongest effects suggest that patients showing high baseline expression of transcripts in one process (≈290) and low expression in a different process (≈ 300) may be less likely to respond to treatment

Summary and way forward

- ***When N << p, one study will not provide strong biomarker evidence*** in realistic scenarios.
  - ❑ Need biomarker data at different doses/exposures in within-arm analysis
  - ❑ Need same inclusion criteria in different arms to tell prognostic from predictive markers
  - ❑ Integrated analysis combining patient-level data across multiple studies

    \+ can increase power

    \- requires investing in collection of biomarker data and metadata from several studies (clinical biomarker database) and data analysis models accounting for between-study variability
- ***Annotations enable interpretation of estimated associations by pathway and process***

\+ Operational characteristics of analysis model enable prospective planning and reporting of RNAseq endpoints

\- When choosing a data analysis model:

- annotation-naïve methods ignore genomic knowledge in analysis of patient-level data
- annotation-based methods rely on genomic knowledge that is incomplete and dynamic

- ***Thank you for your attention***

B/U: Hallmarks of cancer define dimensions of tumorigenesis

Cell, Vol. 100, 57–70, January 7, 2000, Copyright ©2000 by Cell Press

The Hallmarks of Cancer

Douglas Hanahan* and Robert A. Weinberg†
*Department of Biochemistry and Biophysics and Hormone Research Institute
University of California at San Francisco
San Francisco, California 94143
†Whitehead Institute for Biomedical Research and Department of Biology
Massachusetts Institute of Technology
Cambridge, Massachusetts 02142

"We suggest that the vast catalog of cancer cell genotypes is a manifestation of six essential ***alterations in cell physiology*** *that collectively dictate malignant growth"*

Hallmarks of Cancer: The Next Generation

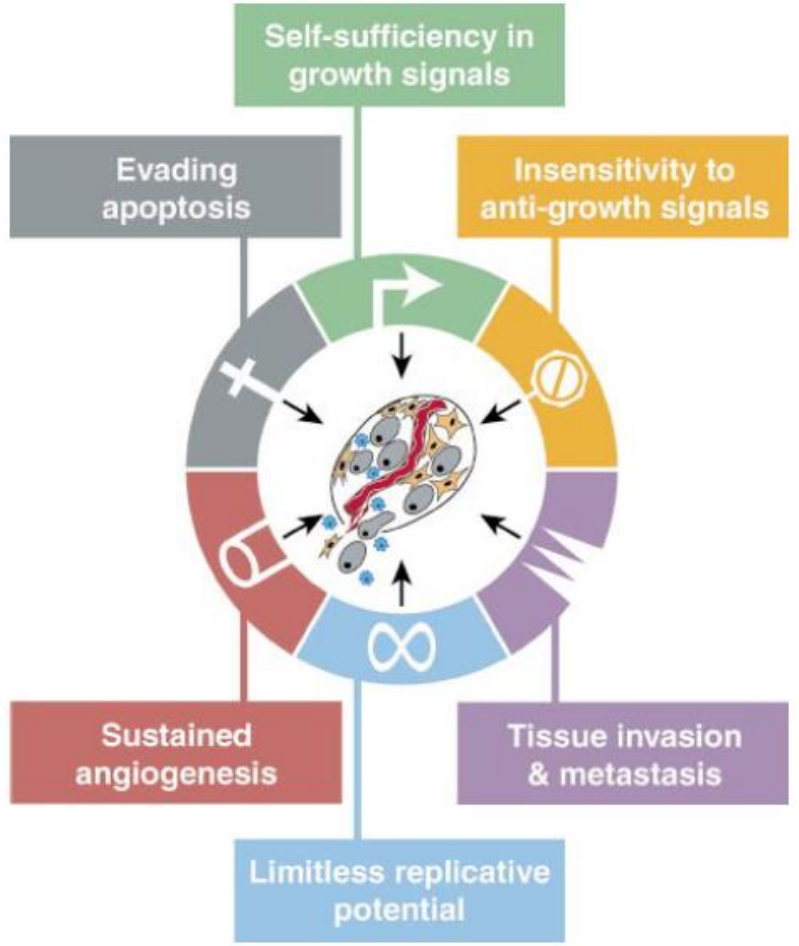
Douglas Hanahan[1,2,*] and Robert A. Weinberg[3,*]
[1]The Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, EPFL, Lausanne CH-1015, Switzerland
[2]The Department of Biochemistry & Biophysics, UCSF, San Francisco, CA 94158, USA
[3]Whitehead Institute for Biomedical Research, Ludwig/MIT Center for Molecular Oncology, and MIT Department of Biology, Cambridge, MA 02142, USA
*Correspondence: dh@epfl.ch (D.H.), weinberg@wi.mit.edu (R.A.W.)
DOI 10.1016/j.cell.2011.02.013

Revisiting the hallmarks of cancer

Yousef Ahmed Fouad[1], Carmen Aanei[2]

[1]Faculty of Medicine, Ain Shams University, Cairo, Egypt; [2]Hematology Laboratory, Pole de Biologie-Pathologie, University Hospital of St Etienne, St Etienne, France

Am J Cancer Res 2017;7(5):1016-1036
www.ajcr.us /ISSN:2156-6976/ajcr0053932



Hanahan et al, Cell, 2000: "*We suggest that […] cancer […] is a manifestation of six essential* ***alterations in cell physiology***"