Experimental design and data life cycle for omics data experiments

A short introduction

Dr. Friederike Ehrhart, Assistant professor

Maastricht University, Department of Bioinformatics – BiGCaT

Maastricht University Medical Centre, Department of Psychiatry and Neuropsychology

Helis Academy

Interreg

EUROPESE UNIE

Vlaanderen-Nederland

Europees Fonds voor Regionale Ontwikkeling

Content

- Research strategies and workflows for omics data generation
- Raw data quality control and preprocessing
- Data life cycle – making data FAIR, especially re-usable

Research strategies

Hypothesis-driven research "reductionistic"

Hypothesis/Theory

Experiment

Falsification

Some hard thinking

New hypothesis

Data driven research "holistic"

Samples

Unbiased omics strategies

Data driven results

Data driven hypothesis/theory

Example: Cancer research

Data – driven: Which pathways/processes are affected in cancer cells?

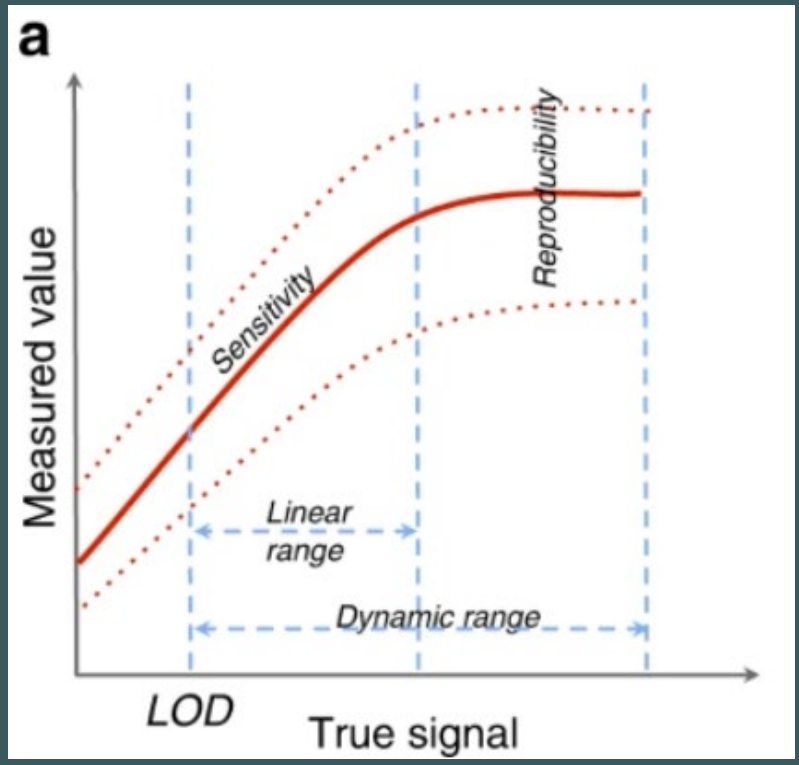
- Experimental design: tumor samples vs. healthy controls
- Collection of tissue samples from tumors and healthy tissue
- Extract molecules of interest (mRNA, microRNA, proteins, metabolites)
- Quantify molecules of interest
- Quality control and statistics

- List of molecules with changed quantities
- Follow up analysis and interpretation

Lauren

Considerations for experimental design

- Sensitivity – reproducibility - coverage
  - RNAseq – all depends on sequencing depth, the more the better – but also more expensive
  - Mass Spectrometry
    - targeted approaches are generally more sensitive than untargeted
    - reproducibility is influenced more by sample preparation protocol
    - coverage is limited in targeted approaches
- Statistical power
  - How to do it properly – check: https://doi.org/10.1093/bioinformatics/bti456
  - What has been accepted in the current research?
    - Transcriptomics - about 20 samples (plus 20 controls)
    - Mass spectrometry/micro arrays – min. 4
    - **Depends on previous studies – check literature and ask your experts for the specific method!**

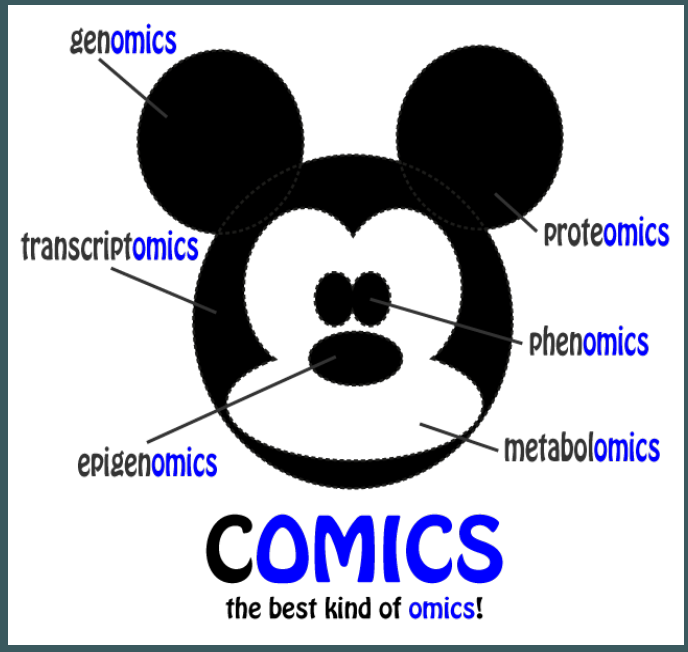


Limit of detection

https://doi.org/10.1038/s41467-020-16937-8

Molecules of interest

- Gene expression
  - Transcriptome
    - mRNA
    - microRNA
    - Whole RNA
  - Proteins
    - Proteome
    - Peptidome
    - Special proteome: e.g. phospho-proteomics
- Metabolite profiles
  - Targeted metabolomics
  - Untargeted metabolomics
  - Special metabolomics - lipidomics

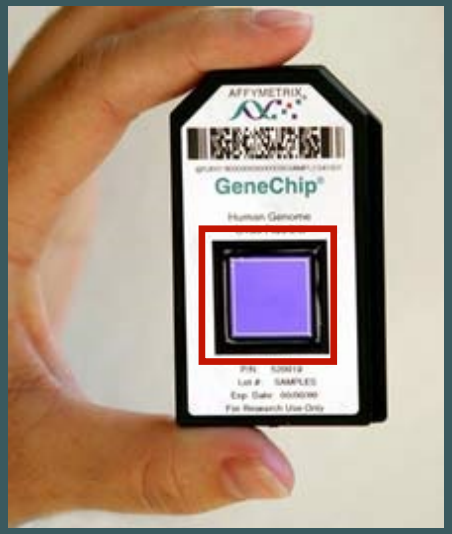


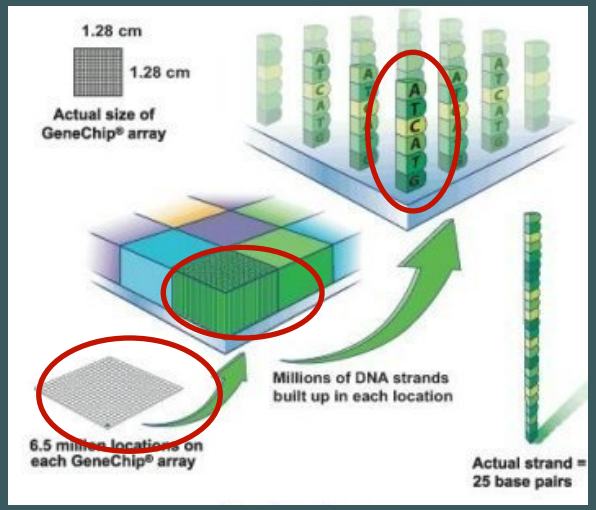
http://www.stathiskanterakis.com/?p=286

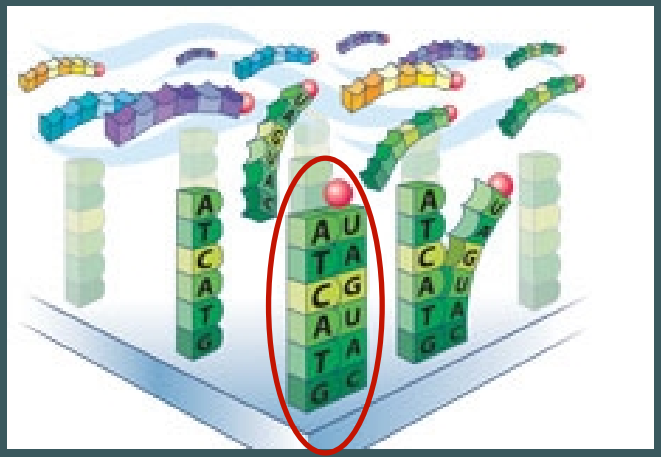
The nature of omics data

Molecules	Method	Kind of data
Transcriptome	Microarray	Fluorescence light intensities
Transcriptome	RNAseq	Counts
Proteome/peptidome	Mass spectrometry	Relative (or absolute) peak size – fold change between experiment and control or absolute values for concentrations
Metabolome		

Affymetrix chips: one sample per array







For Affymetrix chips each gene is measured by dozens of probes that are randomly distributed across the chip; these probes together form a probeset

Affymetrix Chips

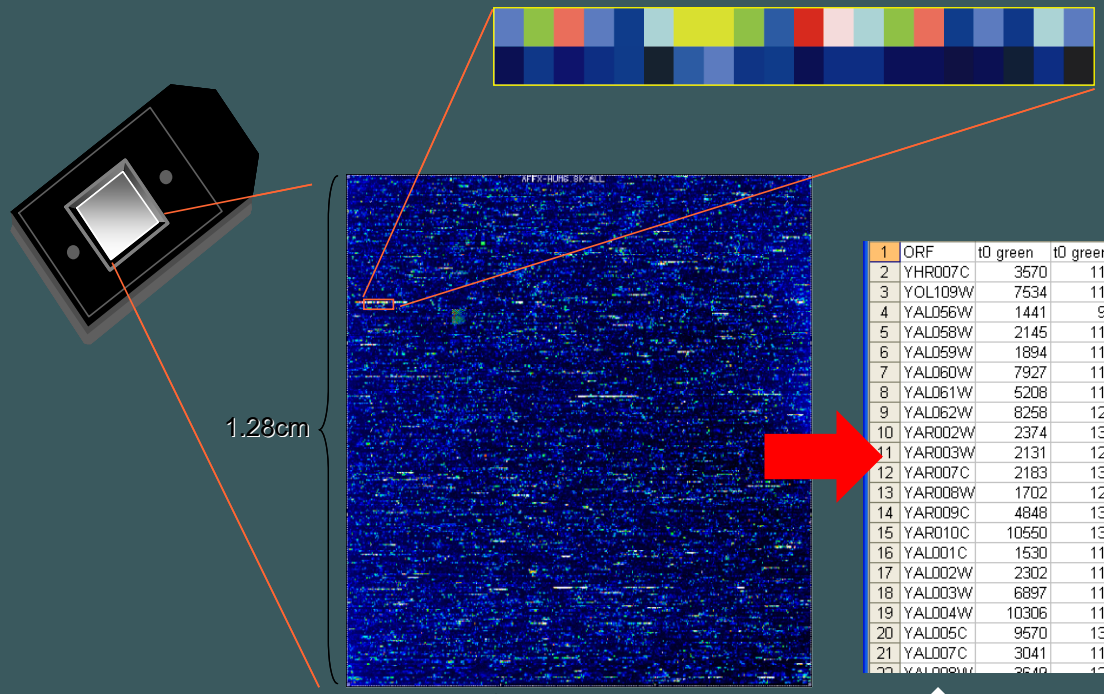


Image of Hybridized Probe Array

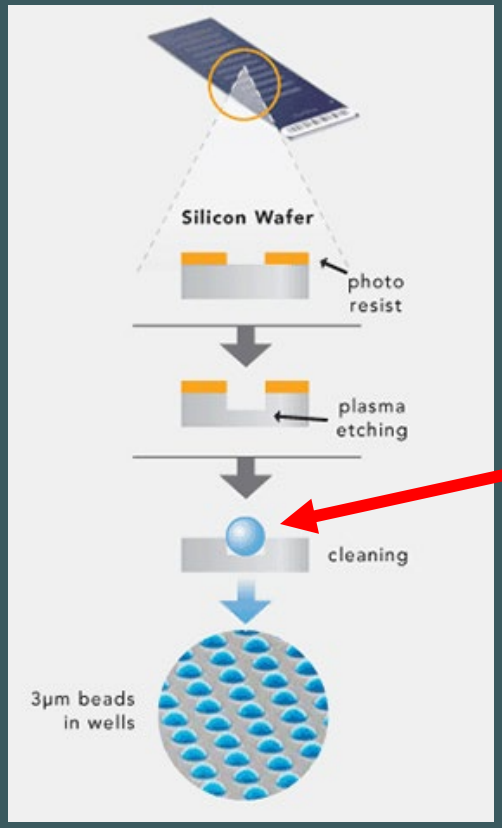
Values

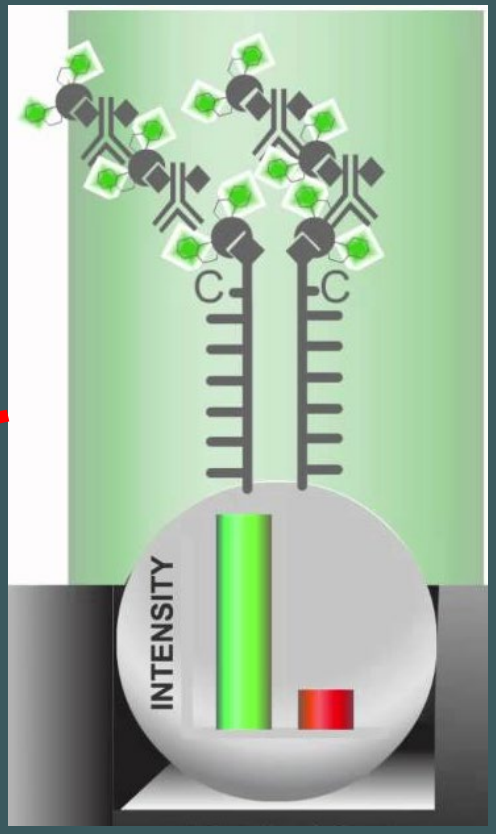
	ORF	t0 green	t0 green bk	t0 red	t0 red bkg	t0.5 green	t0.5 green	t0.5 red	t0.5 red bk	t2 green	t2 green bk	t2 red	t2 red bkg
2	YHR007C	3570	1132	3643	692	3858	1213	5102	1052	2477	1351	3850	785
3	YOL109W	7534	1159	12218	622	7016	1386	5418	576	6119	1470	8272	872
4	YAL056W	1441	996	1043	569	2873	1062	2465	384	1984	1361	1537	858
5	YAL058W	2145	1168	1740	631	2623	1291	1768	670	2122	1535	1486	926
6	YAL059W	1894	1109	1578	575	2145	1052	801	442	1784	1385	1069	789
7	YAL060W	7927	1143	8770	694	9361	1484	5820	772	6740	1586	4029	978
8	YAL061W	5208	1171	5664	756	5914	1108	6008	494	3492	1376	3517	759
9	YAL062W	8258	1224	9527	664	5637	1836	22504	2094	4015	1474	21303	873
10	YAR002W	2374	1308	1838	752	3632	1156	2451	511	2675	1168	1881	643
11	YAR003W	2131	1230	1397	636	2668	1368	2265	580	1848	1184	1652	632
12	YAR007C	2183	1373	1553	794	3170	1179	6450	508	2191	1209	5920	650
13	YAR008W	1702	1214	964	603	2106	1397	1160	590	1635	1250	1743	662
14	YAR009C	4848	1356	4079	748	6508	1277	5457	493	4770	1191	3480	619
15	YAR010C	10550	1361	9306	748	11736	1503	10471	687	9254	1363	7756	742
16	YAL001C	1530	1118	1018	607	2221	1151	1233	421	1818	1407	1171	798
17	YAL002W	2302	1104	1881	614	2705	1493	2307	746	2102	1460	1603	892
18	YAL003W	6897	1160	7621	705	12021	1244	3263	479	6281	1450	2750	762
19	YAL004W	10306	1187	13176	718	12818	1568	8520	804	13036	1506	7086	811
20	YAL005C	9570	1305	13796	857	11039	1308	8848	594	9246	1470	4087	855
21	YAL007C	3041	1142	2768	665	4013	1530	2306	800	2629	1404	2471	834

Probe identifier

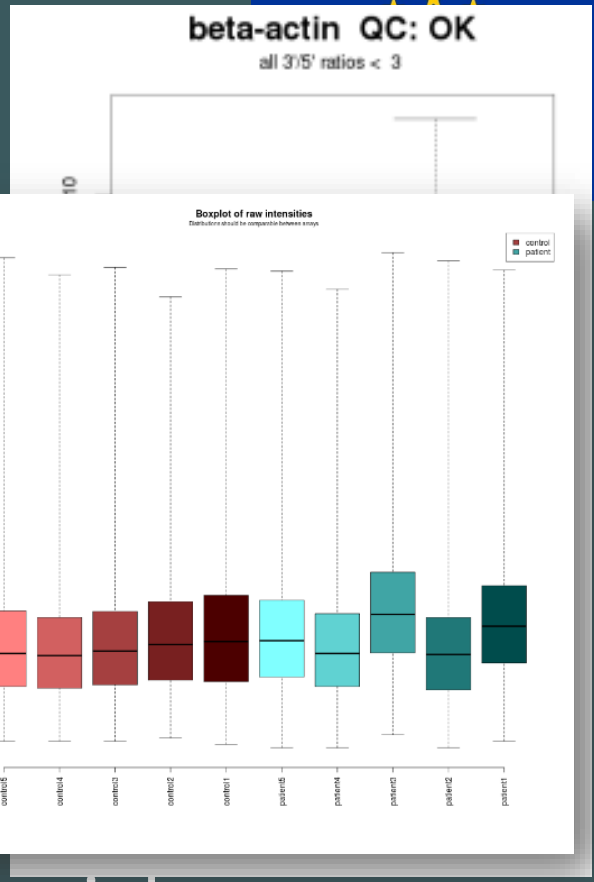
Raw data

Illumina: bead chips





From raw to processed data



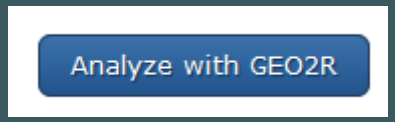
- Quality Control
- Pre-processing
  - Background correction
  - Normalisation
  - Filtering
  - Annotation

https://bio.tools/

- Recommended R packages and tutorial: https://wiki.bits.vib.be/index.php/Analyze_your_own_microarray_data_in_R/Bioconductor

From raw to processed data

- Statistical evaluation
  - T-test
  - Correction for multiple testing
  - ANOVA / modelling
  - > List of differentially expressed genes
- Further analysis
  - Pathway/GO/network analysis



Data publication

- With publication of the paper, also the raw and/or processed data needs to be published
  - ArrayExpress at EBI, Gene Expression Omnibus (GEO) at NCBI
  - MetaboLights for metabolomics/proteomics
  - Zenodo or figshare for more general data
- Standards for proper description for publication of data:
  - MIAME - Minimum information about a microarray experiment
  - MIAPE-MS - Minimum information about a proteomics experiment, MS
- Metadata annotation
  - Standardized language / ontologies
  - MeSH terms for disease descriptions
  - No abbreviations/codes
- Scientific data – journal for data publications

https://fairsharing.org/

Thank you for your attention!
Any questions?

Interreg Vlaanderen-Nederland
Europees Fonds voor Regionale Ontwikkeling
Maastricht University
EUROPEAN JOINT PROGRAMME RARE DISEASES
Helis Academy