Cornell Notes – RNA‑seq & scRNA‑seq in Cancer Genomics (Prof. Jun Wang, 24SEP2025)

Aligned to lecture slides; expanded with contextual notes and primary literature.

# 1) Introduction & NGS Context

|  |  |  |
| --- | --- | --- |
| **Cue (Keywords / Questions)** | **Notes (Detailed explanations)** | **Examples from literature** |
| What problem does RNA‑seq solve? | RNA‑seq profiles the transcriptome (which genes are expressed and at what levels), enabling discovery of differential expression, splicing, fusions, novel transcripts, RNA‑level variants, and co‑expression networks in cancer. Compared with microarrays, RNA‑seq captures a broader dynamic range and novel isoforms without prior probe design. | *The Cancer Genome Atlas (TCGA, 2012) leveraged RNA-seq across thousands of tumours, linking genomic variants with transcriptome alterations in multiple cancer types.* | |
| Bulk vs single‑cell RNA‑seq | Bulk RNA‑seq averages signal across many cells (good for group‑level changes but masks heterogeneity). Single‑cell RNA‑seq (scRNA‑seq) measures each cell separately, resolving intra‑tumour heterogeneity, rare clones, and immune/stromal states. |  | |
| Human Genome Project → NGS | HGP took ~15 years and ~$3B. With NGS, whole genomes now sequenced in ~24 h at ~$1K, making RNA‑seq routine in translational oncology. |  | |

**Summary:** RNA‑seq (bulk) quantifies average expression; scRNA‑seq resolves cell‑to‑cell heterogeneity—critical for cancer where mixed cell states drive outcome.

# 2) NGS Strategies

|  |  |  |
| --- | --- | --- |
| **Cue (Keywords / Questions)** | **Notes (Detailed explanations)** | **Examples from literature** |
| WGS vs WES vs RNA‑seq | WGS: all genomic sequence (3 Gb). WES: coding exons (~1–2%). RNA‑seq: expressed transcripts (mRNA ± ncRNA). Choose WES to focus on protein‑altering variants; RNA‑seq to connect genotype to gene activity and splicing. | *Mardis (2008, Annu Rev Genomics Hum Genet) demonstrated how choosing WGS, WES, or RNA-seq fundamentally alters the biological questions answered — e.g., WES for driver mutation discovery, RNA-seq for transcriptome-level reprogramming.* |

**Summary:** Pick WES for coding mutations; RNA‑seq when expression/splicing/pathways matter; WGS for comprehensive variant landscapes.

# 3) Bulk RNA‑seq – Experimental Design & Library Prep

|  |  |  |  |
| --- | --- | --- | --- |
| **Cue (Keywords / Questions)** | **Notes (Detailed explanations)** | **Examples from literature** | |
| Poly‑A selection vs rRNA depletion | Poly‑A selection enriches mature mRNAs (efficient for coding transcriptomes). rRNA depletion removes rRNA and retains mRNA plus non‑coding RNAs (e.g., lncRNAs), better for degraded RNA or ncRNA studies. | *Zhao et al. (2014, BMC Genomics) showed that polyA+ selection biases against non-coding RNAs and partially degraded RNA, while rRNA depletion captures lncRNAs and degraded samples more comprehensively.* |
| Data generation pipeline (high level) | 1) RNA isolation → 2) Library prep (Poly‑A or rRNA‑depletion) → 3) Sequencing (short reads) → 4) Alignment to genome/transcriptome → 5) Quantification of gene/isoform counts → 6) QC/normalisation → 7) Downstream analyses. |  |
| Visualising alignments | Genome browsers (e.g., IGV) display read coverage across exons and splice junctions to sanity‑check mappings and splicing signatures. |  |

**Summary:** Library choice determines what you see: Poly‑A for coding mRNA; rRNA‑depletion to include lncRNAs. IGV is essential for qualitative checks.

# 4) Bulk RNA‑seq – Core Analyses

|  |  |  |
| --- | --- | --- |
| **Cue (Keywords / Questions)** | **Notes (Detailed explanations)** | **Examples from literature** |
| Alignment & expression quantification | Short reads mapped to reference (transcriptome and/or genome). Expression summarised as counts per gene/isoform (then normalised). |  |
| Differential expression (DE) | Compare tumour vs control (or time points) to find significantly up/downregulated genes (control false discovery rate, e.g., FDR < 0.05). |  |
| PCA & clustering | Principal component analysis and clustering check whether biological groups separate as expected and detect outliers/batch effects. |  |
| Pathway analysis | From DE genes, perform enrichment (DAVID, Ingenuity Pathway Analysis). Alternatively, GSEA ranks all genes to detect subtle, coordinated pathway shifts. EnrichmentMap can visualise pathway communities | *In breast cancer datasets, GSEA identified enrichment of oxidative phosphorylation and MYC targets despite few DEGs, illustrating how pathway-level tools capture coordinated tumour biology that single-gene analyses miss.* (Subramanian 2005) |

**Summary:** DE pinpoints gene‑level shifts; PCA/clustering checks global structure; pathway tools translate gene lists into biology (processes, pathways).

# 5) Cancer Biology Applications

|  |  |  |
| --- | --- | --- |
| **Cue (Keywords / Questions)** | **Notes (Detailed explanations)** | **Examples from literature** |
| Example: estrogen treatment of breast cancer cells | Observed patterns: cell‑cycle programmes down; oxidative phosphorylation up. Suggests metabolic re‑wiring under estrogen signalling. | *Creighton et al. (2009) showed how estrogen signalling alters gene expression in breast cancer, downregulating cell-cycle programmes and upregulating oxidative phosphorylation, linking directly to therapy response.* |
| Why OXPHOS matters | Upregulated OXPHOS indicates increased mitochondrial ATP production and a survival/adaptation route under treatment pressure. |  |

**Summary:** RNA‑seq reveals treatment‑induced programme shifts (e.g., metabolism), informing mechanisms of response and resistance.

# 6) RNA‑level Variants, ASE, Fusions & Splicing

|  |  |  |
| --- | --- | --- |
| **Cue (Keywords / Questions)** | **Notes (Detailed explanations)** | **Examples from literature** |
| Variant & allele‑specific expression (ASE) | RNA‑seq can capture expressed variants and allelic imbalance (bi‑ vs mono‑allelic expression) impacting dosage and function. |  |
| Gene fusions | Chimeric transcripts from genomic rearrangements (e.g., BCR–ABL). Detected by split‑reads/read‑pairs spanning fusion junctions. Tools mentioned: TopHat‑fusion, SOAPfusion. | *BCR–ABL1 in leukaemia and TMPRSS2–ERG in prostate cancer are classic fusions detected by RNA-seq. Both act as oncogenic drivers (Nowell & Hungerford, 1960)\*, while fusions like EML4–ALK in lung cancer highlight their importance as therapeutic targets (Soda, et al. 2007).* |
| Alternative splicing & isoforms | Coverage over exons and splice junctions reveals exon skipping, alternative donors/acceptors, intron retention; isoform‑specific expression can be estimated. | *rMATS and other differential splicing tools can reveal clinically relevant isoform switches, e.g. exon skipping in FGFR2 and CD44 in breast cancer.* (Griffith 2010) |

**Summary:** Beyond expression, RNA‑seq reveals functional variants, allele bias, oncogenic fusions, and isoform re‑wiring relevant to tumour behaviour.

# 7) Co‑expression Networks & lncRNAs

|  |  |  |
| --- | --- | --- |
| **Cue (Keywords / Questions)** | **Notes (Detailed explanations)** | **Examples from literature** |
| Co‑expression | Genes with correlated expression imply shared regulation or pathway membership; network modules can associate with phenotypes. |  |
| Long non‑coding RNAs (lncRNAs) | Typically >200 nt, often lowly expressed and cell‑type specific; participate in DNA/RNA/protein interactions to regulate transcription and chromatin. Ribo‑depletion is preferred to capture non‑polyadenylated lncRNAs. |  |
|  | lncRNA linked to metastasis and poor prognosis across tumours; highly conserved and transcriptionally active. | *MALAT1 promotes metastasis by modulating splicing factors, while HOTAIR recruits PRC2 to silence tumour suppressors. The PCAWG census also pinpointed conserved oncogenic lncRNAs such as LINC00511 and PVT1*. (Yao 2019; Carlevaro-Fita 2020) |

**Summary:** Network‑level views reveal coordinated programmes; lncRNAs (e.g., MALAT1) add a regulatory layer tied to prognosis and metastasis.

# 8) scRNA‑seq – Wet‑lab Workflow & Platforms

|  |  |  |
| --- | --- | --- |
| **Cue (Keywords / Questions)** | **Notes (Detailed explanations)** | **Examples from literature** |
| Plate vs droplet | Plate‑based: cells sorted into wells; high per‑cell depth; lower throughput. Droplet‑based (e.g., 10x): cells barcoded in oil droplets; high throughput (~10k cells/run; ~5–8k usable; ~50k reads/cell). |  |
| General 10x workflow | Encapsulate single cells + barcoded beads → reverse transcription → library prep → sequencing → count matrix per cell (UMIs). |  |

**Summary:** Plate methods suit targeted, deep profiling; droplet methods scale to thousands of cells, enabling atlas‑level tumour microenvironment maps.

# 9) scRNA‑seq – What We Learn in Cancer

|  |  |  |
| --- | --- | --- |
| **Cue (Keywords / Questions)** | **Notes (Detailed explanations)** | **Examples from literature** |
| Tumour microenvironment (TME) | Resolve T/NK/B cells, macrophages, fibroblasts, endothelial and stromal subsets; quantify states (e.g., exhausted vs effector CD8+). |  |
| Tumour heterogeneity & evolution | Identify malignant subclones and lineage/state hierarchies; track differentiation trajectories and potential resistance precursors. | *Lambrechts et al (2018) identified distinct fibroblast subsets (myofibroblastic vs. inflammatory CAFs) in lung cancer; Yeo et al. showed PI3K inhibition redirected luminal progenitors towards resistant basal-like states; Zhang et al. (2021) found strong patient-specific malignant signatures, but a shared enrichment of exhausted CD8+ T cells across tumours.* |
| Cross‑patient patterns | Non‑cancer cells often cluster by cell type across patients; malignant epithelial cells show pronounced inter‑ and intra‑patient heterogeneity. |  |

**Summary:** scRNA‑seq dissects TME composition and malignant cell states, exposing mechanisms of immune evasion and therapeutic resistance.

# 10) scRNA‑seq – Computational Approaches

|  |  |  |
| --- | --- | --- |
| **Cue (Keywords / Questions)** | **Notes (Detailed explanations)** | **Examples from literature** |
| Sub‑clustering & NMF | Unsupervised methods (incl. NMF) discover gene programmes and fine‑grained subpopulations within major lineages (e.g., T‑cell or fibroblast subsets). |  |
| Trajectory (pseudotime) | Orders cells along differentiation/transition paths to infer gene programmes changing en route to resistant or metastatic states. |  |

**Summary:** Clustering/program discovery + trajectories connect discrete cell states into processes, illuminating how tumours adapt.

# 11) Practical Limitations & Integration

|  |  |  |
| --- | --- | --- |
| **Cue (Keywords / Questions)** | **Notes (Detailed explanations)** | **Examples from literature** |
| Challenges | Cost; technical noise (dropout); sample dissociation biases; batch effects; complex analysis pipelines. |  |
| Multi‑omics integration | Combine scRNA‑seq/bulk RNA‑seq with chromatin accessibility, methylation, CNV/mutation and limited proteomics to link regulation → expression → phenotype. |  |

**Summary:** Mind the technical caveats. Integrated multi‑omics offers mechanism‑level insight beyond expression alone.

# 12) Suggested/Referenced Papers

|  |  |
| --- | --- |
| **Cue (Keywords / Questions)** | **Notes (Detailed explanations)** |
| Tumour microenvironment (lung) | Lambrechts D. et al., Nat Med 2018 – phenotype moulding of stromal cells in lung TME; landmark for stromal/immune diversity. |
| Breast cancer heterogeneity (mouse) | Yeo et al., eLife 2020 – distinct patterns of cell‑state heterogeneity and differentiation trajectories in breast cancer models. |
| Esophageal SCC ecosystem | Zhang et al., Nat Commun 2021 – intra‑/inter‑tumour epithelial heterogeneity; immune/stromal contrasts between tumour and normal. |
| Cancer lncRNA census | Carlevaro‑Fita et al., Commun Biol 2020 – conservation and functional evidence of lncRNAs in tumorigenesis (PCAWG). |

**Summary:** Key studies validate scRNA‑seq’s power for TME and heterogeneity; lncRNA resources frame non‑coding drivers in cancer.

# References (selected)

1. **Mardis ER.** Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008;9:387-402.
2. **Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM.** Comparison of RNA-Seq by poly(A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics. 2014;15:419.
3. **Subramanian A, Tamayo P, Mootha VK, et al.** Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-15550.
4. **Creighton CJ, Li X, Landis M, et al.** Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. Proc Natl Acad Sci U S A. 2009;106(33):13820-13825.
5. **Nowell PC, Hungerford DA.** A minute chromosome in human chronic granulocytic leukemia. Science. 1960;132(3438):1497.
6. **Soda M, Choi YL, Enomoto M, et al.** Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. Nature. 2007;448(7153):561-566.
7. **Griffith M, Griffith OL, Mwenifumbo J, et al.** Alternative expression analysis by RNA sequencing. Nat Methods. 2010;7(10):843-847.
8. **Yao RW, Wang Y, Chen LL.** Cellular functions of long noncoding RNAs. Nat Cell Biol. 2019;21(5):542-551.
9. **Carlevaro-Fita J, Lanzós A, Feuerbach L, et al.; PCAWG Consortium.** Cancer LncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. Commun Biol. 2020;3:56.
10. **Lambrechts D, Wauters E, Boeckx B, et al.** Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med. 2018;24(8):1277-1289.
11. **Yeo SK, Zhu X, Okamoto T, et al.** Single-cell RNA-sequencing reveals distinct patterns of cell state heterogeneity in mouse models of breast cancer. eLife. 2020;9:e58810.
12. **Zhang X, Lan Y, Xu J, et al.** Dissecting esophageal squamous-cell carcinoma ecosystem by single-cell transcriptomic analysis. Nat Commun. 2021;12:5291.

\*(first description of the Philadelphia chromosome, later shown to be the BCR–ABL1 fusion)