Contents

Chapter 1

Biology of Cancer

Challenges in cancer genomics [thoughts]

- Identification of epi-drivers. Their mechanisms: non-coding mutations vs. DNA methylation vs. other epigenetic changes. The contribution of epi-drivers to cancer. Targetting pathways.
- Structural variations in cancer: identification of driver genes from events affecting multiple genes?
- Carcinogenesis and cancer progression: how many driver mutations are required? What is the trajectory of clonal expansion in cancer: wait for all mutations, or each driver mutation provides some advantage? (Figure 3, Martincorena & Campell, Science, 2015) What is the relative importance of intrinsic (e.g. cell divisions) vs. environmental factors?
- Metastasis: driver by somatic mutations or epigenetic changes? Result from one special clone in the original tumor or from random tumor cells colonizing a new site?
- Drug target identification (including targets of immuno-therapy): what are good targets? Ex. tumor suppressor genes are often difficult to target.
- Cancer prevention and early detection using knowledge of driver mutations.

Cancer talk by Michael Bishop [iBiology]

- Cancer as a genetic disease: Src discovery in chicken. Myc amplification. Ras point mutations.
- Tumor suppressor genes: often act in recessive way.
- Cancer prognosis: Myc amplification predicts survival. Gene expression signatures: e.g. 70 genes in breast cancer.
- Drug resistance: possible mechanisms are transporters (pump drugs out of cells), mutations in drug target (e.g. resistance to Gleevec via mutation in Bcr-Abl), alternative pathway (e.g. EGFR inhibitor, resistance by mutations in downstream gene Ras).
- Treatment of APL (PML/RAR-alpha fusion): a compound from TCM, and another chemical. Highly effecitive if used in combination. Show that they attack different parts of the target (bimodal attack of the single target).
- Synthetic lethal: in Myc-amplifed tumor, inhibition of CDK1 leads to cell death (in all Myc-cells tested). Remark: possible explanation, Myc sends proliferation signal, while CDK1 inhibitions sends the stop cell cycle signal. The two conflicting signals lead to lethality?

- Other examples of synthetic lethality: (1) Myc: drug that targets chromosome passenger complex (cytokinesis). Early phase, due to apoptosis; late phase, due to Autophagy genes. (2) Between K-Ras and CDK4. (3) BRCA1 and PARP synthetic lethal: the two are important for two DNA repair pathways, homologous recombination and base-excision repair, respectively. Remark: defects in two pathways may create too much DNA damage.

Cancer as microevolutionary process [20.1, MBOC]

- Cancer types by tissue of origin: carcinoma from epithelia cells (80%) - most proliferating cells are epithelia. Sarcoma - connective tissue and muscle.
- Cancer changes growth regulation: contact inhibition, e.g. in petri dish, normal cells form a single layer, but cancer cells can form multiple layers.
- Warburg effect: increase glucose uptake (could be 100 times higher), and most are used for glycolysis, then lactate; only 5% is used for TCA. Possible advantage: carbon provides biosynthetic needs (instead of becoming CO2), side effect of mito. damage (which can trigger apoptosis).
- Survive stress and DNA damage: disable apoptosis. Cancer cells usually die from necrosis (lack of oxygen), rather than apoptosis - its quite common in the center of tumor.
- TME: stroma may facilitate tumor. TME may have: fibroblast, white blood cells, endothelia cells and lymph nodes.
- Overcome replicative senescence: telomerase, or homologous recombination to maintain telemere.
- Invasion and metastasis: (1) Invasion: free from physical tethering, change of adhesion molecules. (2) Enter lymph nodes (clump) or blood (single cells). (3) Metastasis: new colonies (may die).

Cancer critical genes [20.2, MBOC]

- Many ways to activate oncogenes: SNVs, fusion and enhancer hijacking. E.g. Myc: amplification, point mutation, genome rearrangement to put Myc into the enhancer of antibody in B cells. EGFR: loss of extracellular domain.
- Many ways to disrupt tumor-suppressor genes: e.g. Rb, universal regulator of cell cycle. (1) Genetic: need two hits, sometimes one genetic, the other epigenetic, or two different genetic events. (2) Epigenetic silencing: promoter hypermethylation.
- Key pathways repeatedly mutated in cancer: cell growth and cell division are two processes, requiring different signals (mitogens and growth factors). Another common process is p53, stress and DNA damage.
- PI3K/mTOR/Akt pathway: cell growth signal, mTOR activates glucose uptake. PTEN is a tumor suppressor that shuts down this pathway.
- P53 pathway: P53 responds to various stress, including DNA damage, hypoxia, free radicals, sudden oncogenic changes. The responses coordinated by p53 may include: stop cell cycle, senescence, apoptosis. P53 mutations are dominant negative: p53 works as tetramers, so mutation of one copy can disrupt the whole complex.
- Different pathways mutated at different tissues: different signaling processes are involved in proliferation in different tissues. Ex. EGFR in glioblastma, Wnt in gut, TGF-beta in pancreas, NOTCH1 in T-cell lymphoma.
- Mouse studies to define roles of cancer genes: transgenic or knockout mice. Results show that usually multiple mutations are required.

- Tumor heterogeneity: Figure 20-30. A tree of 100 breast cancer cells, 3 subclones, with one having KRAS amplification.
- Metastasis: (1) Invasion: involve EMT. (2) Circulation/travel: angiogenesis, often depend on VEGF. (3) Colonization: only a small percent of cells survive, possibly stem cells.
- Cancer stem cells: only a small percent (e.g. 1/1000) of cells maintain tumor, the rest are transient cells or differentiated ones.
- Colorectal cancer: Stem cells that maintain tissues, replacing all epithelial cells in one week. Starts with polyps. Take decades to become tumor.
- Key mutations in colorectal cancer: APC in 80% (beta-catenin inhibitor), TP53 in 60%. beta-catenin in 5-10% (often a patient has either APC or beta-catenin). DNA mismatch repair genes: a different subtype - mostly SNVs and indels but not chromatin changes. Generally tumors have either genome instability or DNA mismatch repair deficiency.
- Progression and order of mutations: often APC first, then genome instability or mismatch repair, usually p53 later (only when cells enter hyperproliferation, p53 is activated).
- **Lesson**: proliferation signals and pathways vary across tissue of origins. Multiple ways to activate/shut down a pathway, e.g. in colorectal cancer, Wnt pathway can be activated by APC LoF mutation or beta-catenin.

Hallmarks of Cancer: the next generation: Part I [Hanahan and Weinberg, Cell, 2011]

- Sustain proliferation: (1) Signals are often growth factors. Tumors can sustain signaling in several ways: secrete growth factors - autocrine signaling; stimulate normal/stromal cells to send growth signals; or hyperactive signaling. (2) Disruption of negative feedback mechanisms: e.g. PTEN, mTOR (cross-talk with PI3K/Akt).
- Evade growth suppression: (1) Contact inhibition: E-cadherin coupled with intra-cellular signaling via NF2. LKB1: epithelial polarity gene. (2) TGF-beta: anti-proliferative at early stage (tumor suppressor), however in late stage, redirect tumor to EMT (oncogene).
- Resisting cell death: (1) Apoptosis: Bcl2 -— Bax, Bak ¿ apoptosis via cytochrome c. (2) Autophage: similar to apoptosis, a response to stress, and a barrier to tumor. Dual role: early stage is tumor suppressor but later stage and during treatment may enhancer tumor cell survival. (3) Necrosis: trigger inflammation, and release factors, e.g. IL-1, that stimulates proliferation.
- Enabling replicative immortality: (1) Telemere: (2) Excessive proliferative signaling can trigger senescence.
- Angiogenesis: (1) Some tumors: not much angiogenesis, some tumors, very extensive. (2) Could acquire by mutations, VEGF or oncogenic genes, or by Tumor-associate stromal cells, especially innate immune cells such as macrophages. (3) Inhibitors of angiogenesis exist, e.g. TSP-1.
- Invasion: (1) Cell adhesion molecules: E-cad, and N-cad (promotion of cell migration). (2) EMT: central role, controlled by TFs such as Snail, Twist. Activation may be autonomous or stromal cells, e.g. TAM and breast cancer: EGF from TAM to cancer, and CSF1 from breast cancer cells to TAM. (3) Other types of activation may be possible, e.g. stimulate stromal cells to secrete EM-degrading enzymes. (4) EMT changes may be plastic, e.g. when colonizing a new tissue, higher motility is not advantageous.
- Metastasis/colonization: (1) Micrometastases dormancy: can be maintained by nutrient starvation (autophage), immune system, EM. (2) Could happen early in tumor progression. (3) A major question: are metastatic tumor cells endowed with the ability to survive in foreign environment or do they evolve such adaptation? (4) Reseeding can happen.

- **Analysis**: barriers of metastasis may include: (1) Proliferation signals may differ: e.g. ER+ breast cancer cells rely on Estrogen, which is not available in another site. (2) Lack of supportive stroma: tumor often rely on stroma, e.g. TAMs, to provide signals for angiogenesis or invasion. (3) Immune surveillance: intact in new sites. Immune inhibition of a small number of tumor cells may be too weak to overcome immune system.

- **Lesson**: a gene often have pleiotropic functions, affecting multiple aspects related to tumorgenesis. Ex 1. E-cadherin may affect both growth inhibition and invasion. TGF-beta: both proliferation and EMT. Ex 2. Oncogenic signaling (e.g. Ras): also upregulate VEGF signaling, and activates hypoxia response, including glucose import. Ex 3. EMT affects both invasion and cell differentiation.

- **Lesson**: Genetic heterogeneity of cancer: not all hallmarks are present in a tumor, e.g. angiogenesis, and a hallmark of cancer can often achieved by different ways. Ex 1. Growth factor signaling: could be activated by autocrine siganling (production of more growth factors), paracrine signaling (stroma), or hyperactive pathways with normal growth factors. Ex 2. angiogenesis: activation by oncogenes such as Ras and Myc, or inflammatory cells.

- **Lesson**: order of events during cancer progression. Ex. angiogenesis usually early, while resistance to stress later (problem only when excessive oncogenic signaling is turned on).

- **Lesson**: the key challenge is to understand how tumor cells acquire the hallmarks/overcome the barriers. Its important to consider the difficulty of acquiring a hallmark via different means. E.g. several ways to activate invasion, mutation of E-cadherin, or activation of TFs controlling EMT, or activation of signaling process. The last step may be easier to achieve, e.g. by sending a signal to the neighboring stromal or immune cells.

Hallmarks of Cancer: the next generation: Part II [Hanahan and Weinberg, Cell, 2011]

- Enabling characteristic: genome instability. May involve changes to DNA repair system, or sensors and regulators of this system.

- Enabling characteristic: tumor-promoting inflammation, could facilitate tumor in various ways: proliferation signal (e.g. EGF), angiogenesis, invasion (e.g. secretion of matrix degrading enzymes or activation of EMT).

- Emerging hallmark: reprogramming energy metabolism. Facilitate tumor by increasing biosynthesis. Hypoxia response system may be activated in tumor: e.g. glucose transporter overexpression. May exist several subpolulations, some creates lactate as waste, some use lactate as energy source.

- Emerging hallmark: evading immune system.

- Tumor microenvironment (TME): genetic and epigenetic heterogeneity within cancer cells. (1) Cancer stem cells: origin from tissue stem cells or partially differentiated progenitor cells. Difference lies in epigenetics. EMT may be important for stem cell state (self-renewal). Conversion of CSC and non-CSC states may be common: phenotypic plasticity. (2) Genetic heterogeneity: subclones.

- TME: Endothelial cells: Lymphatic vessels may help seed/transport cancer cells.

- TME: Innate immune cells: facilitate tumor because of functions of these cells in tissue repair, wound healing and homeostasis. These cells, including Macrophages, Neutrophils and MDSCs, usually occur transiently during these processes, but they may persist in chronic inflammation.

- Interaction of tumor and TME: metastatic niches, may be due to intrinsic factors of an environment or the effect of circulating factors released by tumor.

- **Challenges of therapy**: often transient response. Explanations: redundant signaling pathways within a hallmark, and shifting from one hallmark to another, e.g. angiogenesis inhibitor induces shift to increased invasion and local metastasis.
- Q: What maintains cancer stem cell states? EMT should be relatively common (required for invasion).
- **Analysis**: what may trigger inflammation and change of TME? Possible causes: (1) Genetic changes in tumor. (2) Stress response, apoptosis or necrosis may trigger inflammation. (3) Pre-existing (chronic) inflammation.

Coming Full CircleFrom Endless Complexity to Simplicity and Back Again [Robert Weinberg, Cell, 2014]

- Experiments that transformed normal cells to tumor: two mutant genes, collaborating with one another [1983]; later experiments, a few genes.
- "At most, one can develop correlations between certain complex data sets (e.g., expression array analyses) and prognosis"
- "The data that we now generate overwhelm our abilities of interpretation, and the attempts of the new discipline of systems biology to address this shortfall have to date produced few insights into cancer biology beyond those revealed by simple, home-grown intuition"
- 'We lack the conceptual paradigms and computational strategies for dealing with this complexity. And equally painful, we dont know how to integrate individual data sets, such as those deriving from cancer genome analyses, with other, equally important data sets, such as proteomics"

Chapter 2

Somatic Mutations in Cancer

Advances in understanding cancer genomes through second-generation sequencing [Meyerson & Getz, NRG, 2010]:

- Heterogeneity of cancer genomic data: a cancer specimen contains a mixture of malignant and non-malignant cells and, therefore, a mixture of cancer and normal genomes (and transcriptomes). Furthermore, the cancers themselves may be highly heterogeneous and composed of different clones that have different genomes.
- Structural variation of cancer genome: in mutation frequency, in global copy number or ploidy, and in genome structure. The analysis of mutations must also be adjusted for the ploidy and the purity of each sample and the copy number at each region.
- WGS in cancer: the main benefits are (1) chr. rearrangement events; (2) alternations in non-coding genome. Examples of translocations in solid tumor: TMPRSS2-ERG in prostate cancer and EML4-ALK in non-small cell lung carcinoma.
- Transcriptome sequencing: mainly for fusion, but can also be used to detect somatic mutations. However, finding a matched normal sample for comparison is a challenge, as normal tissue is unlikely to express exactly the same genes as the tumour sample. E.g. recurrent mutations in (FOXL2) in ovarian cancer.
- Detecting SNPs and small indels in cancer: requires mutation calling in both the tumour DNA and the matched normal DNA, coupled with comparison to a reference genome. Also assessment of statistical signifance of mutation counts: driver mutations.
- Detecting copy number changes: NGS offers advantages such as higher resolution (no saturation), precise delineation of the breakpoints of copy number changes.
- Detecting chr. rearrangements: using both WGS and RNA-seq. WGS is the most costly. Transcriptome sequencing is highly cost efficient but is limited to the detection of coding fusion transcripts.
- Unique computational challenges of cancer genomics:
  - The need to simultaneously analyse data from the tumour and patient-matched normal tissue to identify rare somatic events (for example, somatic single nucleotide variations are 1,000 times less frequent than germline variants)
  - The ability to analyse very different and highly rearranged genomes
  - The ability to handle samples with unknown levels of non-tumour contamination and heterogeneity within the tumour.

- Alignment and assembly: the difficulty of correctly assigning rearranged sequences mean that de novo assembly of cancer genomes, is likely to become the most powerful approach.
- Mutation detection: As somatic genome alterations are rare, any method that detects mutations in cancer must do so with low false-positive rates. Alternatively, a naive somatic mutation caller can be built by independently applying a germline single-sample mutation caller to the tumour and normal data sets; somatic events are those detected in the tumour and not detected in the normal data.
  - A key parameter defined for each mutation is its allelic fraction - the expected fraction of reads in the tumour that harbour the mutation among all reads that map to the same genomic location. The allelic fraction captures the local complexity of the tumour genome, the non-tumour contamination levels and any mutation-dependent experimental or alignment bias, and is also affected by the ploidy of the tumour and the copy number of the region.

Cancer Genome Landscapes [Vogelstein, Science, 2013]:

- Mutation statistics in cancer genomes:
  - In common solid tumors such as those derived from the colon, breast, brain, or pancreas, an average of 33 to 66 genes display subtle somatic mutations (missense or nonsense). About 95% of these mutations are single-base substitutions and the rest indels.
  - Variation of mutation rates across cancer types: Melanomas and lung tumors contain about 200 nonsynonymous mutations per tumor. These larger numbers reflect the involvement of potent mutagens (ultraviolet light and cigarette smoke, respectively). Tumors with defects in DNA repair form another group of outliers (even more). At the other end of the spectrum, pediatric tumors and leukemias harbor far fewer point mutations: on average, 9.6 per tumor.
- Basic mutation model: e.g. colorectal cancer. The first mutation often in APC, then the second one often in KRAS, giving the cells a growth advantage (clonal expansion). Eventually more mutations (PI3K, TP53, etc.), leading to cancer and metastasis. It is estiamted that each driver mutation provides only a small selective growth advantage to the cell.
- Influence on mutations in tumor: tissue origin - some are self-renewing tissues (usually more mutations), age of patients - accumulation of passenger mutations. More than half of the somatic mutations identified in these tumors occur during the preneoplastic phase (growth of normal cells).
- Mutation rates of tumor: applying the molecular clock model, it was found that the rates at which point mutations develop in advanced cancers are similar to those of normal cells. And two unambiguous conclusions. First, it takes decades to develop a full-blown, metastatic cancer (17 years from benign to cancer). Second, virtually all of the mutations in metastatic lesions were already present in a large number of cells in the primary tumors (2 years to metastasize).
- Mutations for metastasis: consistent genetic alterations that distinguish cancers that metastasize from cancers that have not yet metastasized remain to be identified.
  - Hypothesis: epigenetic changes, not yet identified because of heterogeneity. Or no such mutations: a simple model where the released cells randomly settle in a new environment.
- Statistics of chromosome changes:
  - Though the rate of point mutations in tumors is similar to that of normal cells, the rate of chromosomal changes in cancer is elevated. Therefore, most solid tumors display widespread changes in chromosome number (aneuploidy), as well as deletions, inversions, translocations, and other genetic abnormalities.
  - Most solid tumors have dozens of translocations; however, as with point mutations, the majority of translocations appear to be passengers rather than drivers.

  - There are roughly 10 times fewer genes affected by chromosomal changes than by point mutations (Figure 3).
  - Identification of driver genes: difficult with deletions and duplications that affect multiple genes. But easier with translocations (fusion) and gene ampliciations.

- Driver and passenger mutations:
  - Even in a driver gene, not mutations are driver mutations. Ex, in APC, only one nonsense mutation is driver, and the rest missense mutations are passenger mutations.
  - Genes may confer selective advantages on tumor in different ways: somatic mutations, or change of expression through DNA methylation or other changes. Call them Mut-driver and Epi-driver genes, respectively. Epi-driver genes can be identified via experimental manipulation.

- Statistical methods to identify driver genes:
  - Mutation frequency: corrected for gene length, genomic context, etc. However, difficult because the background rates of mutation vary so much among different patients and regions of the genome (and often different in cancer).
  - **Pattern of mutations** (Figure 4): oncogenes are recurrently mutated at the same amino acid positions, whereas tumor suppressor genes are mutated through protein-truncating alterations throughout their length.
  - 20/20 rule: To be classified as an oncogene, we simply require that $> 20\%$ of the recorded mutations in the gene are at recurrent positions and are missense. To be classified as a tumor suppressor gene, we analogously require that $> 20\%$ of the recorded mutations in the gene are inactivating.
  - Application to (COSMIC) database. Though all 20,000 protein-coding genes have been evaluated in the genome-wide sequencing studies of 3284 tumors, with a total of 294,881 mutations reported, only 125 Mut-driver genes, as defined by the 20/20 rule, have been discovered to date (table S2A). Of these, 71 are tumor suppressor genes and 54 are oncogenes.

- Mut-driver genes:
  - We believe that a plateau is being reached, because the same Mut-driver genes keep being rediscovered in different tumor types.
  - New Mut-driver genes found in NGS: proteins that directly regulate chromatin through modification of histones or DNA, mRNA splicing factors, such as SF3B1 and U2AF1, IDH1 and IDH2 in tumor metabolism.
  - Mut-driver genes from chr. rearrangements: adding a few more driver genes (about 10). All fusion genes that have been identified in at least three independent tumors are listed in table S3. The great majority of these translocations are found in liquid tumors (leukemias and lymphomas) or mesenchymal tumors. Other examples in solid tumor: ERG in prostate cancers (70) and ALK in lung cancers.

- The dark matter? What explains tumorgenesis in addition to the relatively small number of driver mutations we found?
  - Classic epidemiologic studies have suggested that solid tumors ordinarily require five to eight hits. However, the number of mutated driver genes is often three to six, but several tumors have only one or two driver gene mutations.
  - Technical limitations: e.g. in a prostate cancer study, only 159 (63%) of the expected 251 driver gene mutations were identified by NGS indicating a false-negative rate of 37%.
  - Difficulty with non-coding sequences.

  - Epi-driver genes: Human tumors contain large numbers of epigenetic changes affecting DNA or chromatin proteins. For example, a recent study of colorectal cancers showed that more than 10% of the protein-coding genes were differentially methylated when compared with normal colorectal epithelial cells.
  - Criteria for distinguishing epigenetic changes that exert a selective growth advantage from those that do not (passenger epigenetic changes) have not yet been formulated.

- **Four types of cancer heterogeneity**: intra-tumor and inter-patient heterogeneity are well-known. In studies that have evaluated intratumoral heterogeneity by genome-wide sequencing, the majority of somatic mutations are present in all tumor cells. The significance of mutations in branches - not clear, but of less clinicial importance (usu. removed). Intra- and intermetastatic heterogeneity among different metastatic lesions of the same patient is of great clinical significance. It is not uncommon for one metastatic lesion to have 20 clonal genetic alterations not shared by other metastases in the same patient. However, the heterogeneity appears largely confined to passenger gene mutations.

- How can the genomic complexity of cancer be reconciled with these clinical observations: cancer cells sometimes respond to drugs targeting specific genes. (1) $>$ 99.9% of the alterations in tumors are immaterial to neoplasia. (2) there appears to be only a limited number of cellular signaling pathways through which a growth advantage can be incurred.

- Core signaling pathways in cancer: 12, divided into three broad processes.
  - Cell fate: Many of the genetic alterations in cancer abrogate the precise balance between differentiation and division, favoring the latter. APC, HH, and NOTCH pathways.
  - Cell survival: allowing cancer cells to survive in limiting nutrient conditions. Mutations of this sort occur, for example, in the EGFR, HER2, FGFR2, PDGFR, TGF-beta-R2, MET, KIT, RAS, RAF, PIK3CA, and PTEN genes. Some of these genes encode receptors for the growth factors themselves, whereas others relay the signal from the growth factor to the interior of the cell, stimulating growth when activated.
  - Genome maintenance: genes whose mutations abrogate these checkpoints, such as TP53 and ATM, are mutated in cancers.

- Implications on common signaling pathways in cancer: inter-patient heterogeneity, e.g. four lung cancer patients may have: activating mutation of EGFR, activating mutation in KRAS (transmit the signal from EGF), inactivating mutation in NF1 (regulator of KRAS) and BRAF (transmit the signal from KRAS to downstream).

- Using knowledge of cancer genomics for treating and preventing cancer:
  - Problem of tumor suppressor genes: most current cancer targets are oncogenes (most drugs are inhibitors). Most effective treatment using multiple targets, including tumor suppressor genes. Idea: use the knowledge of pathways to find the genes affected by tumor suppressor genes, e.g. Inactivation of the tumor suppressor gene PTEN results in activation of the AKT kinase (so AKT can be a drug target).
  - Immuno therapy: use the knowledge of which genes are mutated in cancer to develop therapy, e.g. cancer vaccine.
  - Cancer prevention: early detection based on the knowledge of driver genes.

Computational approaches to identify functional genetic variants in cancer genomes [Gonzalez-Perez, NM, 2013]

- Mutation mapping and annotation: COSMIC, HGMD, OMIM, dbSNP.

- Assessing the functional impact of mutations: common challenge is a curated set of driver and passenger mutations. Tools: TransFic, CHASM, SIFT, PPH2, Mutation Assessor.
- Finding signs of positive selection:
  - A key issue is gene-specific mutation rates: (1) dN/dS, InVEx (introns vs. exons). However, only applicable to cancer with high mutation rates. (2) Covariates such as mutation contex, replication timing and expression. Ex. MuSiC and MutSig.
  - Functional information: OncoDrivFM, ActiveDriver.
- **Challenges**:
  - Prediction of impact on splicing change, which is an important mechanism of driver mutations.
  - Prediction of functional impact: gain vs. loss vs. switch of function.
  - Personalized therapy: prediction of how mutations affect response to therapy; or prediction of driver mutations in an individual.

Somatic mutation in cancer and normal cells [Martincorena & Campbell, Science, 2015]

- Mutational processes in cancer:
  - Exogenous factors: chemicals, UV, ionizing radiation. Endogenous factors: ROS, replication error, aldehydes.
  - Mutational signatures: ex. APOBEC, C>T and C> TG, with the preceding T.
  - Chromothripsis: clusters of tens to hundreds of rearrangements that occur as a single event. Similar to kategis.
- Positive selection on somatic mutations:
  - Frequency of mutations (Figure 2): (1) By tumor types: some have very frequent mutations, e.g. KRAS in 84% of pancretic cancer, APC 69% in colrectal cancer; some not. (2) By genes: generally heterogeneous, only three show consistent effects across tumor types: TP53, PIK3CA, BRAF.
  - Non-coding mutations: mutiple mechanisms, e.g. juxtaposition of enhancers with genes due to rearrangements.
- Progression of normal cells to cancer:
  - Key issue: how cancer cells acquire enough driver mutations. Depends on both mutation rates and clonal expansion.
  - Mutation rates in somatic cells: 2-10 per cell division, order of magnitude higher than germ cells. The number of mutations in somatic tissues not too far from cancer.
  - Selection and clonal expansion: the model (1) no clonal expansion until all driver mutations occur; (2) each driver mutation carries a small advantage. In sun-exposed skin cells: a large fraction of cell already obtain driver mutations.
- Cancer and aging:
  - Cancer as an example of aging: accumutation of somatic mutations, amplified by clonal expansion, over time lead to reduction of the ability of tissues to function normally.
  - Evolution of protection mechanisms againast cancer: high-fidelity replication, cellular senescence, stem cell hierarchiy, TS genes, immune survelliance, microenv. control.

Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational Prediction of Driver Missense Mutations (CHASM) [Carter & Karchin, CR, 2009]:

- Motivation: mutational frequency alone is not a reliable measure of driver mutations, as rare mutations are known to drive cancer. Ex. IDH1 mutation. Thus design a method that predicts whether a mutation is likely to drive cancer, using the information of the sequence and gene, but not frequency.
- Supervised learning: create a positive (driver) and negative (passanger) class of mutations. Experts determine what genes are oncogenes, and the mutations in these genes are considered driver mutations (2,000 mutations). Negative set: first select genes with at least one mutation; then sample from mutational spectra. Learning by Random Forest (better than SVM).
- Features: 50 features (1) about AA changedestablization, protein localization, position of mutation (binding or active sites), sequence conservation. (2) protein functional annotation: molecular function (kinase, DNA-binding.) (3) Exon-level constraint: average cross-species conservation by PhastCons, SNP density.
- Importance of negative set: previous methods use ns SNP with high MAF as negative set. Show that synthetic mutations have different features (PCA).
- Results: outperform SIFT and PolyPhenl in cross validation; and in two genes, p53 and EGFR. The most predictive features are: exon conservation (#1), COSMIC subst. frequency (#2), PTM (e.g kinase), DNA-binding domain and SNP density.

Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers Fran [Supek & Lehner, Cell, 2014]

- Background: natural selection acting on syn. sites, possible mechanisms: speed or accuracy of mRNA translation, mRNA folding and splicing, translational pausing.
- Syn. mutations are elevated in oncogenes but not tumor suppressor (TS) genes: employing multiple controls. (1) Control genes that match mutational covariates; (2) Introns (InVEx). (3) Neighboring genes.
- Syn. mutations in oncogenes: often in evolutionarily conserved sites, and form spatial clustering (within 5 codons), similar to missense clustering in both oncogenes and TS genes.
- Function of syn. mutations: (1) gain or loss of exon splicing motifs: enrichment of syn mutations within 30nt of exon boundaries; (2) experimental evidence.
- Dosage-sensitive oncogenes (amplication) have an excess of mutations in 3'UTRs, but not recurrently missense-activated oncogenes.
- **Lessons**:
  - Synonymous mutations can be driver mutations.
  - Splicing changes are important mechanism of cancer. Multiple mechanisms: splice site changes (LoF), synonmoyus and missense mutations near exon boundaries.

Discovery and saturation analysis of cancer genes across 21 tumour types [Lawrence & Getz, Nature, 2014]

- Goal: increasing power by combining multiple tumor types; power analysis.
- Data: 5000 tumor samples in 21 tumor types.
- MutSig: CV: mutational rate; FN: average phyloP scores of mutations; CL: % mutations within hotspots (3bp). Combine 3 p-values using Fisher's exact test.

- Gene findings: (1) Single gene analysis: 224 genes. Some tumor types, e.g. READ (rect) and COAD (colon); lung and head-neck. (2) Combine data of all tumor types: 110 genes with 30 new genes.
- RHT (restricted hypothesis testing) analysis: for each tumor type, test only genes that are significant in the remaining tumor types. RHT adds 0 - 13 genes.
- Functions of new cancer genes: hallmarks of cancer, proliferation, genome stability, chromatin regulation, evasion of immune attacks.
- Power analysis: need 600-5300 samples to find genes mutated in 2% or more patients.
- Lessons:
  - Tumor heterogeneity: if simply combine all tumor types, the power will be reduced (comparing with union of results from single tumors).
  - Validation: hallmarks of cancer (related processes), Copy number amplification.

Landscape of somatic mutations in 560 breast cancer whole-genome sequences [Nik-Zainal & Stratton, Nature, 2016]

- Method for detecting driver genes: basic model. 192 cell types, and the relative rates constant across all genes. Each gene has a specific background rate t. Additionally, each type of mutation is associated with a RR: e.g. ω_{nonsense}. The model can be used for single-gene analysis, which is effectively dN/dS test (that treats t as nuisance). However, the power is limited.
- Model of gene-specific rates: Negative Binomial regression of t on covariates, including replication timing, etc. Use negative binomial to include overdispersion (Gamma prior of t).
- Model of non-coding mutations: multiple noncoding classes, including core promoter, UTR, intronic sequences in exon boundaries, enhancers (from ENCODE), ultra-conserved sequences. Each class is modeled separately. The difference of the model is the covariates in the NB regression include local mutation rates (neighboring non-coding mutations). Instead of dN/dS, estimate the expected number of mutations (or rate), and do Negative Binomial test.
- SNV and indel finding: 5 new breast cancer genes, including MED23.
- Structural mutations: in-frame gene fusions not common, mostly in hypermutable regions. Recurrent rearrangements were found to disrupt cancer genes.
- Incorporating copy number changes including homozygous deletions and amplifications, generate a total of 93 probable cancer genes.
- Recurrent mutations in non-coding regions: 5 significant elements, 2 lncRNAs and 3 promoters. In two promoters, recurrent mutations that match mutational signatures, suggesting likely due to mutations rather than driver mutations.
- To assess the contribution of mutation signatures: (1) Extract mutation context, and compare with known signature. The flanking sequences can be important, e.g. TGAACA core, if flanked by ¿9bp paliondrome, has about 200 times higher mutation rates. (2) Extract samples of the mutations of interest, and asses the contribution of mutation signatures in these samples and see if any signature dominate (enriched), comparing with other samples.
- Charactering novel cancer genes (not found in breast cancer): (1) Compare the role of genes: e.g. previously reported as dominantly acting oncogenes, and now there is evidence of TS genes. Use Cancer Gene Consensus as resources. (2) Reported role in other cancer (e.g. MLLT4 in hematological cancer).

- Remark: the paper use a different mutation model: model gene-specific rate using regression (i.e. covariates modeled at the level of gene-specific rate rather per base level).
- Lessons
  - For non-coding mutations, the local mutation rates/density better covariates than replication timing and others.
  - Recurrent mutations could be caused by mutation signatures. Some signatures could have very large effects on mutation rates. The strategies of assess the contribution of mutation signatures.

Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics [Li Ding and TCGA, Cell, 2018]

- Assessing trans-effects of driver mutations: use Moonlight [Moonlight: a tool for biological interpretation and driver genes discovery], test association of mutations of a gene (group into LoF and missense) and expression (DE) of pre-defined gene sets.
- Trans-effects often vary with tumor types for the same gene; also in some cases, the effects of a gene were found only for LoF or missense mutation (Figure 4H). Ex. GATA3 LOF, but not missense mutations (known to have oncogenic function) associated with cell mobility, cytoskeleton.
- OncoIMPACT analysis (patient-specific change of expression): take mutation, expression and PPI network data. For each patient, estimate if an oncogenic process (based on network modules) is dysregulated, and find the associated mutations (patient-specific mutation). Then test association of dysregulated process and mutations across patients.
- Mutual exclusive pattern of driver mutations: across patients (so different patients have different mutations to achieve the same transcriptome changes), e.g. for cell proliferation, KRAS, EGFR, PTEN, NOTCH1; for cell death, TP53, PIK3CA, KRAS, APC, PTEN.
- Somatic mutations and TME (Figure 7): define 6 types of TME based on transcriptome signatures (clustering samples), wound healing, IFN-gamma dominant, inflammatory (C3), lymphocyte depleted (C4), immunologically quiet and TGF-beta dominant. For each type, estimate cell fractions (MPhages, Cytotoxic T, DCs). Association of driver mutations with cell fractions within each of 6 subtypes: e.g. BRAF and NRAS associates with CD8 T cells in C3. In C4: association of EGFR with neutrophil and FGFR3 with macrophage.
- Method for association of drivers and immune cell fractions: using regression (adjusting for total mutations and tissue of origin). Test only 300 driver genes, with filters: only LoF mutations or missense classified by two programs as oncogenic, driver events in at least 10 samples.
- Remark: in Figure 7, the immune modulators are known, and here their effects on immune cells are confirmed with correlation (of expression and cell fractions).
- Remark: OncoIMAPCT analysis relies on patient-specific mutations for a dysregulated process. There are statistical problems of (1) inferring the dysregulated process in only one sample; and (2) assign mutation to that process (using cross-sample information).
- Lesson: convergence of transcriptome changes or processes in cancer, despite many possible mutations. Ex. cell proliferation and cell death: a single mutation is sufficient to drive the change.

The long tail of oncogenic drivers in prostate cancer [NG, 2018]

- Data: 1000 WES of prostate cancer.
- Found 90 Significantly mutated genes. Many unknown. Many genes have low frequency, $< 3\%$ patients.

- Pathway analysis: epigenetic genes, splicing, Wnt signaling.
- Comparison of primary vs. metastatic tumors (Figure 4): genes enriched in metastatic tumors. Also found significantly more frequently altered in metastatic than in primary tumors: Wnt, PI3K, etc.

2.1 Calling Somatic Mutations in Cancer

Somatic mutation calling from DNA and RNA-seq: errors and strategies [personal notes]

- We consider both bulk sequencing and single-cell sequencing (data are pooled). We consider DNA-seq first, then RNA-seq.
- Germline variants: filtering by germline variant database and AF cutoff. However, even if we filter 95
- Misalignment: indel realignment (if using a tool other than GATK), and visual inspection.
- Sequencing error: the model should account for it.
- DNA contamination: mostly should be germline, so germline filter can help.
- DNA damage and PCR errors: if we have single-cell data, check if the mutations occur in multiple cells.
- Allele bias: can change the AF, and a result, germline variants called as somatic.
- In RNA-seq samples, we have additional sources of errors. ASE and RME: both shift the AFs of germline variants. RNA editing: filter with known editing sites.
- CNVs (cancer): can change AFs. Infer CNV events, and adjust for the different ploidy when calling mutations in those regions.
- Analysis: the biggest challenge in scRNA-seq analysis without normal is to distinguish germline (but AF changes due to ASE or MAE) from somatic. Idea: if germline, every cell has the mutations. So if there are some cells with sufficient coverage, but never express both alleles, it rejects germline. We can formulate this as a likelihood model.

Allele-specific copy number analysis of tumors (ASCAT) [Van Loo and Kristensen, PNAS, 2010]

- Model: consider only germline SNPs, assume all copy number changes are clonal. Let ρ be tumor purity, and Ψ_t be average tumor ploidy (a single number for the entire tumor). For a genomic location i, let $n_{A,i}, n_{B,i}$ be the copy number of A and B alleles in tumor (1, 1, for normal tissues). We can then relate logR and BAF with the allele specific copy numbers as:

$$r_i = \gamma \log_2 \frac{2(1-\rho) + \rho(n_{A,i} + n_{B,i})}{2(1-\rho) + \rho\Psi_t} \qquad b_i = \frac{1 - \rho + \rho n_{B,i}}{2(1-\rho) + \rho(n_{A,i} + n_{B,i})} \tag{2.1}$$

  Remark: log-R is defined as the ratio of intensity at a location with genome average, so the denominator use average tumor ploidy.

- Estimation procedure: enumerate a grid of ρ and Ψ_t, and for each grid value, find the best integer values of $n_{A,i}$ and $n_{A,i}$, and then compute a goodness of fit by summing over the errors over all SNPs (where errors are defined by the best fit allele-specific copy numbers from the equation and the integer numbers).
- Application to breast cancer: validation of tumor ploidy by FISH. Different distributions of ploidy and purity in five breast cancer subtypes.

- Remark: The model does not consider subclonal CNVs.

Absolute quantification of somatic DNA alterations in human cancer [Carter and Getz, NBT, 2012]

- Inference of absolute number of somatic CNAs (SCNAs): the observed copy ratio depends on (1) tumor purity: some are normal cells; (2) ploidy of cancer cells; (3) subclonal evolution: possible that some subclones have a SCNA, while others not.
- Input data: relative copy number data. The paper uses homolog-specific copy ratio (HSCR) data, distinguishing paternal and maternal copies. This is done by first phasing the genome using germline SNPs; then estimate the ratio with segmental estimates of allelic copy ratio.
- Relating relative copy ratio and ploidy: Eq. (1), let α be the tumor purity, $q(x)$ be the ploidy of position x, then we have the relative ratio $R(x) \propto \alpha q(x) + 2(1-\alpha)$. This can be normalized. The key parameters are α and mean ploidy of the genome D (normalization constant).
- Model: consider N regions, with relative copy ratio x_i for region i. Also have standard error σ_i and the genomic fraction w_i (the fraction of genome that have the relative copy ratio). We model the underlying ploidy state s_i as multinomial, with mixture weight dependent on w_i (prior). s_i is mostly discrete, but we allow subclonal events (a state). x_i is related to s_i by mixture of normal: if s_i is subclonal, x_i follows uniform distribution; otherwise, the mean of x_i is given by the equation above. ML estimation of purity α and the mean ploidy (the mean of Gaussian distribution).
- Ambiguity of α and D: coupled. Can be resolved by karotype data.
- Use copy number information to infer multiplicity of somatic point mutations: this is related to subclonal evolution. If there is only clonal events, ploidy would determine the multiplicity of SNVs. However, if SCNA happens before a SNV, and there are subclones, then the SNV frequency can be lower than expected based on SCNA (only a fraction of cells have the SNV). The knowledge of purity and ploidy of SCNA can resolve this.
- Note: different orders of SNVs and CNVs imply different allelic SNV ratios.
- Remark: (1) Separate estimation of relative allelic copy ratio and ploidy: is this optimal? (2) Assumption that when subclonal, the HSCR is uniformly distributed?

TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data [Ha and Shah, GR, 2014]

- Model idea: log-R and BAF depends on tumor cell fractions (TCFs) of events. The TCFs form a distinct set of clusters, with spatial continuity.
- Model of log-R and BAF: we have T hetero. germline SNPs, let $l_{1:T}$ be the log-R (normalized by GC content and mapping bias over 1kb bins) and $a_{1:T}$ be the BAFs. Suppose SNP t belongs to cluster z with event g, then the mean of log-R is denoted as $\mu_{g,z}$, and the expected BAF as $\omega_{g,z}$.

$$l_{1:T} \sim N(\mu_{g,z}, \sigma_g^2) \qquad a_{1:T} \sim Bin(N_{1:T}, \omega_{g,z}) \tag{2.2}$$

  where $N_{1:T}$ is read depth. The values of $\mu_{g,z}$ and $\omega_{g,z}$ are determined by tumor purity n, average ploidy ϕ, event g and TCF s_z (Figure 2D, Equations 1-2 in Suppl).
- Spatial model of g and z: 2-factor HMM, modeling transitions of both events g and subclones/TCFs z.
- Inference: EM algorithm to estimate parameters. Search for number of clusters from 1 to 5.

Global copy number profiling of cancer genomes [Wang and Zhang, Bioinfo, 2016]

- LRR model: LRR is the log ratio of read depth between tumor and normal samples in any segment (assuming segmentation has already been done). Let ϕ and ρ be the purity and average ploidy of tumor. Let n_T be the total copy number in a segment, then the expected LRR is given by (copy number in this segment, normalized by average copy number across all segments):

$$\mu^{(r)} = \log_2 \frac{n_T\phi + 2(1-\phi)}{\rho\phi + 2(1-\phi)} + \text{const} \tag{2.3}$$

  where the constant adjust for library size difference between tumor and normal. The actual data of LRR is modeled as normal distribution with variance shared by segments.

- BAF model: consider a heterozygous site, let n_B be the number of B allele copies and n_C be the total copies of two alleles. The expected BAF is then given by:

$$\mu^{(b)} = \frac{n_B\phi + (1-\phi)}{n_C\phi + (1-\phi)} \tag{2.4}$$

  The actual BAF can deviate from this because of mapping bias and other errors. Use another normal model. Note that n_B here is related to n_T in the LRR model, but depends on how B allele is defined.

- Joint model and parameter estimation: assume r_k and b_k (local LRR and BAF) are independent. The parameters are local copy numbers $\mathbf{x}$ and ϕ, ρ. To fit the parameters, use "canonical point" methods: given ϕ, ρ, the LRR and BAF would fall into several discrete values - canonical points (depending on the underlying copy numbers). We then assign each segment into its nearest canonical point, and minimize the total distance over grids of ϕ, ρ.

- Remark: does not have a prior for local copy numbers.

2.2 Mutational Processes in Cancer

DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes [Woo & Li, NC, 2012]

- Data: 22 WGS and a few hundred WES data from TCGA.

- Mutational frequency and covariates: 1Mb region, for each region, x_i covariates including: replication timing, recombination rate, distance to telomere, GC content; and y_i: log2-transformed SNVs. Do ANOVA and find that Replication timing explains a very large fraction of variation of mutation rates in cancer cells, but much less in germline cells.

- Correlation of mutational rate and replication timing: divide all neutral genomic regions into 10 bins by their replication timing, then correlation of replication timing with mutations. Very strong correlation, $R^2 > 0.9$ for cancer (increase of 200% in late replicating regions), but much less so for germ line (only 28% increase).

- Selection in cancer somatic mutations via dN/dS test: In germline, dN/dS is significantly less than 1 for missense mutations. In cancer, close to 1 for both missense and LoF.

- Discussion: why relaxed negative selection or increased positive selection in cancer? Cost of complexity: benefical mutations at the level of organism much rarer; also deleterious effects of somatic mutations are restricted to only mutated cells and their progeny.

- Remark: different analysis strategies to study correlation of covaraites and mutations. (1) Correlation of covariates in many regions with mutations; (2) Divide all sequences into bins by their covariates, then correlate with mutations. In (2), ignore all variations within a bin, thus much higher correlation (also depend on interval size - larger intervals, higher correlations).

Signatures of mutational processes in human cancer [Alexandrov, Nature, 2013], Deciphering Signatures of Mutational Processes Operative in Human Cancer [Alexandrov, Cell Reports, 2012]

- Mutation signature model: let K be the number of mutational types (e.g. 6), G be the number of samples. Our data is the $K \times G$ matrix M. We assume there are N signatures, each signature is represented by a vector of probabilities (proportion of each mutation type) $p_k, 1 \leq k \leq K$. And define E, $N \times G$ matrix, as the percent contribution of signatures to each sample. Then we have:

$$M = P \times E \tag{2.5}$$

  as the MOM estimator of P and E. This is NMF, and can be solved by minimizing $\|M - PE\|^2$ through multiplicative update. The method does not have statistical uncertainty, so we use bootstrapping to obtain $\bar{P}$.

- Mutational signatures: 21 in 30 tumor types. Patterns: variable even in the same cancer type.

- Interpreting mutation signatures: some are known, others can be correlated with epidemological and cancer features. Ex. signature 13: ds break, APOBEC signature 2.

- Kategis: localized hypermutation, clusters of C>T and/or C>G mutations, often preceded by T. Suggest an underlying role of APOBEC.

2.3 Methods for Somatic Mutations

Statistical analysis of pathogenicity of somatic mutations in cancer [Greenman, Genetics, 2006]

- Model: consider k mutational types, $1 \leq k \leq 6$, let l_k, m_k, n_k be the number of silent, missense and nonsense mutations, respectively. Let L_k, M_k, N_k be the length (number of bps) of three types of mutation, and ρ_k be the mutation rate parameter. Our model is:

$$l_k \sim \text{Pois}(L_k \rho_k) \qquad m_k \sim \text{Pois}(M_k \rho_k \phi_k) \qquad n_k \sim \text{Pois}(N_k \rho_k \psi_k) \tag{2.6}$$

  where ϕ_k and ψ_k represent selection strength. Specifically, $\phi_k = \eta M_k^c / M_k + (1 - M_k^C / M_k)$, where η is the relative risk of missense mutations. We have only one parameter ϕ_k since the RR of causal mutations and the proportion of causal mutations are coupled.

- Testing: our goal is to test if $\phi_k = 1$ or $\psi_k = 1$. While we can test more general hypothesis, in practice, we can test against $\phi_k = \phi \neq 1, \forall k$. This is done using conditional test to remove the nuisance parameter ρ_k. Developed LRT.

- Functional domains: develop a test for domain-specific selection (one domain has higher selection than the others).

- Results in breast cancer: 25 breast cancer samples and sequence of 518 kinase genes. Found average selection about 1.4 for missense (not significant) and 5 for nonsense and splicing.

- Remark: the conditional test would lose power. This problem is fixed in newer versions [Stephens, Nature, 2012].

The landscape of cancer genes and mutational processes in breast cancer [Stephens & Stratton, Nature, 2012]

- Model: let $l_{kg}^i, m_{kg}^i, n_{kg}^i$ be the number of silent, missense and nonsense mutations of mutational type k in gene g of individual i, and L_{kg}, M_{kg}, N_{kg} be the corresponding length. We model these counts as mixture model: there are two driver genes groups, those driven by missense and by nonsense mutations, with proportions α and β respectively. The selection strength are λ and μ respectively for the groups of driver genes.

- Inference: use EM. Once obtain parameters, using Bayes Theorem to test if a gene is a driver gene.
- Results: 100 breast cancer tumors, most ER positive. About 70 SNVs per sample and 2-3 indels. Found 9 new cancer genes.

Mutational heterogeneity in cancer and the search for new cancer-associated genes (MutSigCV) [Lawrence & Getz, Nature, 2013]

- Motivation: most of the driver genes found by existing methods are false positives. Ex. 178 WES from lung cancer, found 450 genes at $q < 0.1$. About 1/4 are Olfactory receptors, many are large genes $> 4,000$ AAs.
- Mutational heterogeneity: if genes have very different mutation rates, even if one uses the correct average rate (of all genes), there will be false positive finding of genes.
- Data for characterizing mutation heterogeneity: 3,000 WES/WGS samples in 27 tumor types. Average of 4.0 non-silent coding mutations / Mb per sample.
- Comparison within and across tumor types: could vary by 1,000 fold. Pedaitric cancers: as low as 0.1/Mb, while melanoma and lung cancer could be as high as 100/Mb. Even within a single tumor type, thee rate can vary dramatically, e.g. AML, 0.01 to 10/Mb.
- Variation of mutational types: 96 mutational types. Different cancers are enriched with different types. NMF analysis: Reduce the dimensionality (96), with each spectrum represented as a linear combination of six basic spectra. Ex. Lung cancer: C $\rightarrow$ A mutations.
- Regional variation of rates: strong correlation with replication timing (late replication, higher rates, perhaps due to lack of free nucleotides), and with gene expression level (low expression, higher rate, due to TCR). These two features explain most of spurious findings.
- MutSigCV mutation model: correct for mutation rates.
  - Gene level mutation rate: to incorporate covarates (replication timing, expression and Hi-C), for each gene, obtain its nearest 50 neighors with these covariates (some distance cutoff is applied). Remove the neighbors with very different mutation rates. Then average.
  - Patient-specific rate for each mutational type: multiply the gene mutation rate by scaling constants for patient and mutational type, respectively.
- MutSigCV testing of genes: (1) for each gene, each mutational type and each patient, test its mutational count against rate using Beta-Binomial model, and obtain p-values. Nonsense mutations are weighted more (add an arbitrary constant). (2) Combine the 6 mutational categories S_{gp} (for gene g and patient p): roughly top 2 p-values. (3) Combine the info of all samples: sum of S_{gp} across p.
- Application of MutSigCV to lung cancer study: from 450 to 11 genes found.
- Remark/Question:
  - The paper says that heterogenity across samples may also produce false positive findings, why?
  - Mutation model: analogous to hierarhical model (using group average), but very heuristic. And a potential of bias: remove genes with very different mutation rates.
  - Testing: combining 6 mutational categories: only top 2 contribute? Obtaining null distribution of test statistic: permutation, but may be possible to obtain closed form using MGF (sum of RVs).

DrGaP: A Powerful Tool for Identifying Driver Genes and Pathways in Cancer Sequencing Studies [Hua & Lu, AJHG, 2013]

- Poisson model of somatic mutations: the general idea is to consider the number of mutations in each type (11 in total): in the driver gene, these numbers would be higher than expected. Let n_{ijk} be the number of non-silent mutations in sample i, type j and gene k, and m_{ijk} be the number of silent mutations. Let η_{ij} be the background rate in sample i, type j, and α_{jk} be the additional rate due to the driver gene property. Then:
$$m_{ijk} \sim \text{Pois}(\eta_{ij} M_{jk}) \tag{2.7}$$
where M_{jk} is the number of positions in type j of gene k (the exposure) for silent mutations. And similarly:
$$n_{ijk} \sim \text{Pois}((\eta_{ij} + \alpha_{jk}) N_{jk}) \tag{2.8}$$
where N_{jk} is the same number (exposure) for nonsilent mutations. This model allows one to test the hypothesis: $\alpha_{jk} = 0, \forall j$ using LRT.

- Estimating background mutation rates: the MLE from the likelihood model, using silent mutations. To avoid the problem of 0-count, use a prior of Beta distribution.

- Incorporating sequencing coverage: when defining the exposure in Poisson model, consider only regions with sufficient coverage $>$ 8x.

- Application to lung cancer: high mutation rates, about 6 mutations/Mb. Using a sample of $>$ 100 WES, identify 110 driver genes at 5% FDR.

Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome [Davoli and Elledge, Cell, 2013]

- Classification of TSG and OGs: collect 71 known TSGs and 54 OGs, and neutral genes, classify the two groups. Features measure dN/dS ratio, dLoF/dS ratio, etc and trait the models on Pan-Cancer data. Use Lasso to select for features.

- Top features are: (1) TSG: LoF/Benign, HiFi missense/Benign and splicing/benign, deletion frequency (2) OG: entropy for missense (spatial distribution), HiFi missense/benign, amplification frequency.

- TUSON Explorer: ignore the CNV features, and derive a TSG and OG score for each gene. Either combine p-values or Lasso.

- Genetic architecture of cancer: (1) Estimate number of TSGs and OGs from p-value distributions, about 300 each. (2) Tumor type specificity: relatively few genes found in each tumor type separately, and most genes at FDR ¡ 0.25 were highly ranked in Pan-Cancer results.

- Evaluating new pathways and genes: many well known pathways, new pathways include immune system, esp. antigen processing and presentation, and negative regulation of cell adhesion. New genes in known pathways: discuss a few in DNA damage response.

- Explaining pattern of SCNA using new driver gene list: show amplifcation/deletion of chr. arms can be predicted from the density of TSGs and OGs.

Cancer Evolution Is Associated with Pervasive Positive Selection on Globally Expressed Genes [Ostrow & Hershberg, PLG, 2014]

- dN/dS test for cancer: number of nonsyn and syn. mutations in a test group vs. numbers of these mutations in all genes. Comparison with mutation rate based approach: local mutation rates. Possible bias in this test: not correct for mutational types.

- Average dN/dS in all somatic mutations of breast cancer: 0.82 vs. 0.28 in germline. Also use dMF/dLF (more functional vs. less functional) from PPH and SIFT: similar patterns.

- Explanation of relaxed selection or increased positive selection in cancer vs. germline. (1) Limited in tissues and cell types; (2) Hitchhiking: reduce effect of negative selection; (3) Smaller population size.
- Results of dN/dS in all genes: found 100 genes with dN/dS significantly greater than 1, most unknow. Among 400 or so known cancer genes, only 4% show significant evidence of $dN/dS > 1$.
- Positive selection in globally expressed genes: they are generally under very strong negative selection, enriched with housekeeping functions.
  - dN/dS in these genes 0.91 vs. non-globally expressed genes, 0.78.
  - Known cancer genes are more likely to be expressed in many tissues.

The dJ/dS Ratio Test Reveals Hundreds of Novel Putative Cancer Drivers [Chen & He, MBE, 2015]

- dJ/dS test: J mutations are defined as mutations in splice sites: the first and last two bases of an intron. Obtain the expected rates of synonymous and splice mutations using all genes. To test one gene, obtain its counts J and S, then use Binomial test (two-sample Poisson test) using expected J and S mutation rates.
- Results of applying the test to 22 tumor types: one tumor type a time. Found 393 putative cancer driver genes at $q < 0.1$, of which 62 are CGC genes, 5.6 fold enrichment. Table 1: some genes have very strong evidience, eg. RPS27 has 30 J but 0 S mutations.

MADGiC: a model-based approach for identifying driver genes in cancer [Keegan Korthauer, Bioinformatics, 2015]

- Frequency-based methods: MutSigCV, MuSic, based on baseline mutation rate (BMR). BMR depends on GC content, sample (e.g. one may have mutation in DNA repair), replication timing (early replication, lower mutation rate), expression level (transcription-coupled repair). Remark: most driver genes habor suprisingly few mutations.
- Functional impact based method: bias of mutations towards high impact ones, Oncodrive FM.
- Positional clustering methods: Oncodriv CLUST, mutations tend to cluster at certain regions. However, the clustering pattern is not obvious even with a dataset of 500 samples from TCGA overian. Use Catelog of Somatic Mutations in Cancer (COSMIC) across thousands of samples to see cluster patterns.
- Model (MADGIC): combine the benefits of these approaches. Let X be a binary indicator of mutational status of a gene in a sample (the gene contains mutation or not in this sample), S be the impact score, and Z a binary variable (to be determined) of the driver gene status. Define a generative model of X and S from Z: $P(S, X|Z)$, the idea is that driver genes have more mutations and higher impact scores.
- BMR model: sample-specific model, incorporating the relevant factors such as replication timing and expression level.
- Results: 463 overian cancer samples (WES), median of 60 mutations. MutSigCV identifies 5 genes, two of which are known drivers. MADGIC finds 19 genes, 58% are driver genes.
- Remark:
  - Functional impacts: different for oncogenes and tumor suppressor genes, thus should be modeled differently.
  - Spatial clustering is not modeled.

Evaluating the evaluation of cancer driver genes [Tokheim and Karchin, PNAS, 2016]

- A new method: 20/20+, use ratiometric features, dN/dS ratio, functional impact bias and positional clustering (entropy). Then derive p-value for each feature, and train and random forest classifier. Use the same 71 TSGs and 54 OGs.
- Evaluation by CGC overlap: TUSON, 20/20+ and MutSigCV show 40
- Other evaluations (1) p-value distribution, (2) method consensus and (3) consistency (random two-way split of samples). Similar patterns.
- Evaluation by number of significant genes: 2020+, TUSON and MutSigCV predicted 150-240 genes, while other five predicted 400-2600 genes.
- Specific tumor types: choose 4 types with different mutation rates. TUSON always predicts fewest genes.
- Simulation to assess the problem of unexplained background variation: significant number of false positives for both mutation rate-based and ratiometric based methods at large mutation counts and high unexplained variability. Ratiometric methods generally less sensitive to unexplained variability.

A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence (DISCOVER) [Canisius and Wessels, GB, 2016]

- Problem: given data, the number of mutations (0/1) of each gene in each sample, test independence of a pair of genes. FET assumes constant mutation rates, and is susceptible to FPs.
- Model: estimate mutation probability p_{ij} for gene i in in sample j. Let X be the number of tumors with mutations in both genes i and k. Then we have: $E(X) = \sum_j p_{ij}p_{kj}$. This leads to a binomial test of X. If p_{ij} are given for each i and j, the distribution of X is Poisson-binomial.
- Testing mutex. in a gene group: use no. tumors with ≥ 2 gene mutations as test statistic.
- Application to each tumor type separately: only find one significant co-occurrence, but several hundred mutex.

Cancer driver gene discovery through an integrative genomics approach in a non-parametric Bayesian framework (iDriver) [Yang & Li, Bioinfo, 2017]

- Model idea: each gene is represented by a feature vector. Fit the background distribution. A driver gene should be far from the center of the background. Features: mutations, gene expression, CNA, DNA methylation and protein. Background distribution: Mixture model with Dirichlet prior. Scoring a gene: Assess the distance of the gene vector to the mean of the background distribution.
- Results: four cancer types, 2 main clusters in each case. One minor cluster, about 4%, is enriched with driver genes.
- Comparison with MutSigCV: show higher enrichment of known cancer genes in the top predictions (FDR cutoff).
- Evaluations: in predicted genes, remove known ones (1) Distance to known driver genes in PPI network (HumanNet). (2) Higher connectivity in PPI network. (3) Strong purifying selection. (4) Higher expression level. (5) Enrichment of cancer pathways in KEGG. Also show how the genes are related to known genes in the pathways. (6) shRNA screening in cell lines: enrichment of genes important for cancer cell survival.
- Remark: the scoring function is strange, in the background model, one cluster actually corresponds to driver genes. This is not used in scoring.

- Remark: comparison with MutSigCV not fair, because the method predicts fewer genes (the top genes are generally better).

Conditional Selection of Genomic Alterations Dictates Cancer Evolution and Oncogenic Dependencies (SELECT) [Mina and Cirello, Cancer Cell, 2017]

- Background: classic example of mutual exclusion (mutex), RAS and BRAF.
- Define candidate events: a few hundred recurrent genes and CNVs.
- SELECT algorithm: (1) weighted MI, where weights correct for mutation frequencies (otherwise, lower weights to rare mutations). (2) Testing statistical significance by permutations. (3) To penalize indirect, transitive effects: correction, discount events often correlated with other events.
- Results: apply to all TCGA data together, and find 400 mutex motifs and 200 co-occurence motifs.
- Some examples: Figure 3B-D. K-RAS: mutex with other MAPK genes. TP53: mutex with P16 (cell cycle) - P16 mutation provide large advantage without P53 mutation. RNF43: WNT signaling, co-occurrence with CD58 (mediate T cell cytotoxicity to tumor). Explanation: CD58 can sustain WNT signaling, however, when there is a activating mutation in WNT (RNF43), Wnt function of CD58 becomes less important, and it would be more advantageous to have LOF of CD58 to escape from immunity.
- Module level analysis: the largest module has ERK and Wnt signaling genes, with both positive and negative interactions.
- Emergence of motifs through conditional selection: motifs are much more common with conditional selection.
- Lesson: mutex may often occur within pathways/modules. Cooccurence: when one mutation "rescues the deleterious effect of another mutation.
- Remark: for such analysis, important to correct for transitive relations.

Universal Patterns of Selection in Cancer and Somatic Tissues (dNdScv) [Martincorena and Campell, Cell, 2017]

- Model: dN/dS test, number of silent and non-silent mutations modeled as Poisson. Silent: the per base rate t_j for gene j, and use substitution model for different kinds of mutations (relative rate). Non-silent (missense, LoF): multiply t_j by ω (one value for each category).
- Modeling t_j: Unif, constant for all genes; Loc: one value for each gene; CV: incorporating covariates. Let t_j follows Gamma distribution, and the number of silent mutations follow NB distribution. Do NB regression: silent mutations vs. covariates. The model estimates the overdispersion parameter and the expected value of t_j based on covariates. This leads to a Gamma posterior for t_j, but the method uses MLE.
- Remark: the model is similar to ours, if we ignore covariates that contributes to expected mutation rates. After fitting the NB regression, the method estimates t_j. The difference is: t_j is estimated via MLE, while we integrate out t_j. This likely makes substantial difference when we have significant uncertainty of t_j.
- Testing driver genes: test if $\omega = 1$, do 2 or 3-degree of freedom test. t_j is estimated by MLE (close form).

- Testing indels: remove all known cancer genes, then fit a NB regression model of number of indels in a gene with covariates. This leads to the estimate of expected indel rate per gene and overdispersion parameter. The p-value from the indel test is combined with that from dNdS with Fishers method. Consider two definitions of indels: (1) number of indels per gene (2) number of unique indel sites: motivated by repeated indels due to artifacts and mutation hotspots. Found that (2) is much better.
- Detecting genes at positive selection: found a small number of false positives due to sequencing artifacts, using a tool to remove them. Idea: the sequencing artifacts and calling errors would be similar in frequency in normal tissues. A black list of 49 genes.
- Overall pattern of dNdS: close to 1 in both tumor and somatic tissues.
- Identification of genes under positive selection: 179 cancer driver genes, about 54% are in CGC.
- Negative selection is largely absent: the observed distribution of dN/dS across all genes is similar to neutral expectation. Fit a mixture model to estimate the proportion of positive and negative selection: less than 1 coding substitutes are lost by negative selection.
- Testing negative selection in individual genes and pathways: no signal gene passing threshold. Many gene sets: essential genes, intolerant genes (from germline), only nonsense mutations in essential genes in the region of single copy show significantly lower dN/dS.
- Possible explanations of lack of negative selection in cancer and somatic cells: buffering effect of two copies, small population size, Mullers ratchet (hitchinghiking of deleterious mutations).
- Estimate the number of driver mutations per tumor sample: 4-10 in all genes across all tumor types. Indels contribute about 0.7 per sample and silent mutations about 0.1.
- Remark: the dndscv test is not easily generalizable, to include more mutation categories, e.g. severe mutations, the extra dof would reduce power.
- Remark: alternative explanation of lack of negative selection: the effect of deleterious mutations (could be many) in individual clones are offset by the positive selection on driver mutations?

Identifying Epistasis in Cancer Genomes: A Delicate Affair [van de Harr and Ideker, Cell, 2019]

- Why not adjusting mutation load can lead to false epistasis? Ex. CMS1 subtype of CRC has much higher mutation rates. P53 mutations are less frequent in CMS1 subtype. This leads to mutex between p53 and MUC16 (large gene): no P53 mutation ¿ tend to be CMS1 ¿ tend to have MUC16 mutations.
- Show that many genes have high associations with total mutation count (high MLA). Most epistasis would occur with low MLA genes. And this can be explained partially by tumor subtypes.
- Challenge (Figure 4): mutations can create tumor subtypes, which may affect mutational process/rate, which affect mutation and epistasis.

2.4 Non-coding Somatic Mutations

Genome-wide analysis of noncoding regulatory mutations in cancer [Weinhold & Lee, NG, 2014]

- Motivation: using a large collection of cancer NGS data to identify the non-coding sequences that drive cancer.
- Data: 863 WGS from TCGA, about 1,000 to 50K somatic mutations per sample. To identify potential enhancers, use 66.9K enhancer-to-gene associations (27K enhancers) from FANTOM.

- Overall mutation rates in different non-coding categories: similar rates in promoter, UTR, coding and enhancers. The rates are higher in introns and even higher in intergenic regions. The results suggest negative selection against mutations that reduce cancer growth or viability.
- Hot-spot analysis: merge all mutations within 50bp. To calculate p-value of each cluster, use negative binomial distribution, taking into account the length and the background mutation rate of each cancer sample (average mutation rate across the whole genome). Results: top cluster is TERT promoter, which contains 2 highly recurrent mutations. The next: PLEKHS1 promoter, 23 mtuations over 20 samples.
- Mutational recurrence analysis: use binomial test for any region (larger than hot-spot analysis, e.g. around 2K). The baseline rate is determiend by either local analysis (10kb flanking region) or global analysis (match the type of sequences against the same type of sequences in the genome, stratified by replication timing). Example: WDR74, 36 non-coding mutations in a 1kb window near its promoter.
- TF motif analysis: evaluate the disruption and creation of new TFBSs by somatic mutations. TERT promoter: the hotspot mutations create novel binding sites for ETS TFs.
- Lessons: somatic mutational landscape is a consequence of both negative and positive selection.

Promoter capture Hi-C-based identification of recurrent noncoding mutations in colorectal cancer [Orlando and Houlston, NG, 2018]

- Data: PHi-C in colrectral cancer cell lines. Overall mutation rates in PHi-C regions: lower than other parts, could be due to negative selection or chromatin accessibility.
- Testing for driver mutations (separate analysis on MSI or non-MSI cancers): three criteria, clustered mutations, excess of mutations and change of gene expression in mutation vs. non-mutation samples. A single sequence (2-3K) interacting with ETV1 promoter.
- **Lesson**: account for mutation rate confounders in non-coding SNV analysis, including MSI, chromatin accessibility/HiC.

2.5 Structural and Copy Number Variations

Big data mining yields novel insights on cancer [Jiang & Liu, NG, 2015]

- PCA on large cancer expression profile datasets: correct for batch effect, GC bias, etc.
- Inferring somatic copy number alteration (SCNA) from transcriptome data: show that the PCs of neighboring genes are similar.
- **Lessons**: the latent variables, e.g. PCs, may correlate with something of biological interest. Ex. SCNA, cell growth, TF regulation.

Integration of Genomic Data Enables Selective Discovery of Breast Cancer Drivers (Helios) [Sanchez-Garcia & Pe'er, Cell, 2015]

- Background:
  - In breast cancer, only six genes have point mutations in $> 5\%$ of samples, while 87 SCNA (somatic copy number alternation) regions pass this threshold.
  - Over 70% of 140 recurrently altered regions did not contain a known oncogene or tumor suppressor.
  - Limitations of shRNA screen (growth effect of gene knockdown): noisy results, off-target effects, not a good model of tumor (no microenvironment).

- Limitations of existing approach for CNV:
  - Most SCNA detection algorithms, including GISTIC2, compute a null distribution across the entire genome to estimate the significance of alterations. However, the alteration rate can strongly differ across different genomic regions, due to features such as DNA secondary structure and DNA hypomethylation.
  - Found that a number of oncogenes such as BCL2 were not found by GISTIC2 (most popular).
- Identifying significantly altered region (ISAR): ISAR accounts for the local alternation rates. Given a gene (marker) m, define its G-score as the sum of copy number over all samples where the score is higher than a threshold:

$$G(m) = \sum_{i=1}^{N} CN(m,i) I(CN(m,i) > \theta) \tag{2.9}$$

  where $CN(m,i)$ is the copy number of marker m in sample i.
  - Null distribution: use a local sliding window to determine the null distribution of each window, then obtain q-values. If a gene is associated with multiple overlapping windows, use the least significant q-value.
  - Window size: try multiple ones, and use the most siginificant q-value among the window sizes.
  - To define regions: merge multiple consective markers
  - Results: found 87 new regions (vs. 30 reported from TCGA), with average size of 14 genes.
- Idea of integrating gene features and CNV data: the goal is to learn the driver genes from SCNAs by the features of genes. To learn the features, it is not good to use known driver genes - highly biased towards kinases and other genes. The idea is to use the ISAR scores defined earlier to train the features: these scores (at the gene level) are similar to "partially labeled" samples. Gene features: point mutations (MutSig), expression, shRNA screen scores
- Helios algorithm: our data is the SCNA scores of genes, we are trying to infer the latent variables T (driver or not). The prior of T would be the gene features, denoted as X. The model:

$$P(SCNA) = \sum_{t=0}^{1} P(SCNA|T=t)P(T=t|X) \tag{2.10}$$

  The first part: first normalize G scores from ISAR (not comparable across regions), then model the normalized scores as a mixture of exponential distribution, with parameters λ_0 and λ_1 for passanger and driver genes. The features X would include, for example, the MutSig score (log-transformed). The distribution $P(T|X)$ is often based on logistic regression. The extension here is to use Bayesian network of additional layers: (1) each node (input feature) is connected to a node representing sigmoid function; (2) some features are combined: e.g. two shRNA features. The model has parameters not only for the weights of features, but also thresholds.
- Inference: the parameters are λ_t and W (parameters of the Bayesian network) and T are latent variables. Using EM to estimate T and the two sets of parameters. Initialization: start by assigning most confident genes (from SCNA) as positive genes.
- Experimental validation: in top 17 regions, choose the top-scoring genes, 12 in total. Experiment: attachment-dependent growth is a key feature of epithelial oncogenes. Overexpression of 10/12 genes in non-transformed cells lead to increased attachment-independent growth.
- Investigation of one gene (RSF1) in depth: In vitro model: similar to above. In vivo: non-transformed cells overexpressing RSF1, then put in SCID mice, and found primary tumor growth.

- **Lesson**: more complex, Bayesian network prior for binary latent variable.
- Remark:
  - In ISAR scoring, when a gene overlaps with multiple windows, uses the least significant q-value. This leads to loss of information.
  - The method relies on the fact that SCNA data alone has strong information for at least a significant subset of genes, this allows one to learn the features, and then use these weights to refine the labels.

Copy number alterations unmasked as enhancer hijackers [NG, 2017]

- Topological mechanisms that can lead to activation of oncogenes: enhancer in proximity with oncogenes from translocation, deletion and inversion; enhancer duplication; deletion of insulators; insulator spanning tandem duplications.
- Computational analysis: the boundary of rearrangement, and test the recurrence of boundaries.
- Remark: we could test how often a gene is activated, combining all possible mechanisms that leads to its activation. The challenge is to have a predictive model of gene activation.

2.6 Gene Fusion

Bcr-Abl background:

- Translocation: exchange of genetic materials between two regions. Could be unbalanced (gaining or missing genes).
- Philadelphia chromosome: translocation, in which parts of two chromosomes, 9 and 22, swap places. Fusion gene of Abl in chr 9 and Bcr in chr 22. This is a reciprocal translocation, creating an elongated chromosome 9 (der 9), and a truncated chromosome 22 (the Philadelphia chromosome).
- Depending on the precise location of the fusion the molecular weight of the protein can range from 185 to 210 kDa. Three clinically important variants are the p190, p210 and p230 isoforms.
- BCR-Abl transcript is also translated into a tyrosine kinase, constitutively active. Since ABL activates a number of cell cycle-controlling proteins and enzymes, the result of the BCR-Abl fusion is to speed up cell division.

A summary of Bcr-Abl fusion:

- Alternative translocation points: provide alternative Bcr-Abl transcripts across different samples. Two main types reported: (1) break point in Bcr intron 13 and 14, creating p210 form, mainly in CML. (2) Break point in Bcr intron 1, creating p190 form, mainly in ALL, more aggresive.
- Alternative splicing of fusion transcripts: provide alternative Bcr-Abl transcripts in the same sample. Three main types reported: (1) Bcr-2 or 3 fused with Abl-II. (2) Bcr-1 or 14 fused with Abl-II. (3) Bcr-1,13, or 14 fused with Abl-II or IV.

Alternative splicing of RNAs transcribed from the human abl gene and from the bcr-abl fused gene. [Cell 1986]

- Abl: 3' two thirds of sequence in a single exon. Abl has two major isoforms with exonIa and exonIb (from different promoters). The two alternative exon I correspond to exons 1 and 4 respectively of mice.

- Fused transcript is spliced into a 8kb chimera mRNA unique to CML, which contains all Bcr exons up to the translocation point and all Abl exons except exon 1. Translation of this chimera mRNA leads to 210kD protein.
- Breakpoints: variable positions of Abl exon II on most CML patients, on two small regions (several kb) in Bcl at chr. 22 - between exon 2 and 3 or between exon 3 and 4. Results: two fusion transcript (1) Bcr exon 2-Abl exon II - (L-6 junction); (2) Bcr exon 3-Abl exon II (K-28 junction).
- Alternative splicing of the fusion transcript: the Abl exon II has the ability to splice over long distance (exon Ia and Ib), thus it can frequently skip bcr exon 3. In 21 CML samples, 18 has at least one of the two junctions: 4 contained L-6 junction alone, 6 K-28 junction alone and 8 contained both junctions.

Exon-skipping in BCR/ABL is induced by ABL exon 2 [Biochem J, 2000]:

- Background: Alternative splicing of fusion transcripts associated with cancer has been reported in desmoplastic small round cell tumour [35], T-cell acute lymphoblastic leukaemia [36,37] and acute myeloid leukaemia [38,39]
- Alternative break points: the translocation may occur at different regions of Bcr, and produce different fusion proteins(always fused to Abl exon 2).
  - If the break point lies in Bcr intron 13 and 14 region (M-Bcr), the result is a p210 protein with most of Bcr exons up to exon 13/14. In CML, most translocations involve the M-bcr of BCR.
  - If the break point lies in Bcr intron 1 (m-Bcr), the result is a p190 protein with the first Bcr exons. Often in ALL, and more transforming potential.
- Alternative splicing:
  - Production of p210 form: can involve excision of greater than 200 kb of pre-mRNA, including ABL exons 1a and 1b.
  - AS of p210: Bcr exon 1a or 1b. ABL exon 2 can also alternatively splice toBCR exon 13 by skipping exon 14. Co-expression of both the p210- and p190-encoding BCR-ABL transcripts in CML patient samples are reported (from AS of the same fused genome).
- Model of AS of p210 and p190 transcripts: (Figure 6) in the p210 form, the Abl exon 2 promotes splicing that remove the exons from 2-8 and 9-14. The result is Bcr exon 1 joined with Abl exon 2 (p190 form).

Alternative BCR/ABL splice variants in Philadelphia chromosome-positive leukemias result in novel tumor-specific fusion proteins that may represent potential targets for immunotherapy approaches [Cancer Res, 2007]:

- We could detect BCR/ABL transcripts with junctions between BCR exon 1, 13, or 14 and ABL exon 4 in approximately 80% of patients and 84% of cell lines, beside the main fusion transcripts.

Chapter 3

Genetics and Environment in Cancer

3.1 Germline Variants in Cancer

A Polymorphic p53 Response Element in KIT Ligand Influences Cancer Risk and Has Undergone Natural Selection [Cell, 2013]

- SNP in a CRE of KITLG gene: disrupt binding of P53, and change gene expression. KITLG: increase cell proliferation.
- The SNP is under positive selection. Most other SNPs in TP53 response elements are under negative selection.

Shared heritability and functional enrichment across six solid cancers [Jiang and Lindstrm, NC, 2019]

- Heritability estimation: 3 (ovarian) - 25% (prostate). Breast, colorectal, headneck, lung in the middle.
- Genetic correlation between cancer: breast and ovarian; lung and headneck, colecteral.
- Genetic correlation with non-cancer traits (Figure 3): smoking and lung cancer. Years of education and mental disorders (SCZ, depression) vs. breast, lung, head-neck and colorectal. WHR and BMI and lung, head-neck and breast. However, not found BP and cancer. Immune diseases: IBD and eczema, suggestive evidence with lung and breast.
- MR: strong evidence, SCZ > breast, SLE > prostate, age at menopause > breast.
- Enrichment analysis with S-LDSC: epigenome data from 220 tissues. Figure S2: immune strongest in lung and ovarian; and in meta-analysis across 6 tumor types, many immune annotations highly enriched.
- Discussion: the MR result of SCZ/education on cancer risk, not clear if it is mediated entirely by behavior/smoking.

3.2 Germline-Somatic Interactions

Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer [Carter and Ideker, Cancer Disc, 2017]

- Background: Heritability of cancer overall 33%. In TCGA, 10% of patients have rare germline trucations in 100 driver genes. Examples: in JAK2 and EGFR, haplotypes of the gene influence cancer somatic mutations (cis).

- Germline variants and tumor site of origin: first obtain candidate associations at $p < 10^{-5}$. Then empirical FDR control with random associations at FDR < 0.25.
- Analysis: GWAS case-control study of patients, where cases and controls are defined by the status of somatic driver mutations. Ex. somatic MYC amplification vs. non-amplification. Data: 4000 subjects, full genotypes (close to 1M SNPs)
- Interaction between germline and somatic mutations: 62 associations at FDR ¡ 0.25 (test on pairs of 138 driver genes and all markers). No cis-associations. Multiple testing correction: (1) first obtain candidate associations from FET, then re-calibrate p-values from permutations; (2) In validation data, do permutation and estimate the number of associations above a threshold to get empirical FDR.
- RBFOX1 SNPS promotes SF3B1 mutation: SNP has large effects (8 fold increase of SF3B1 mutation). Using RNA-seq data: confirm expression change of RBFOX1 SNP. Then show that in RBFOX1 SNP alt. allele, SF3B1 mutation has a larger effect on splicing. Possible model: two splicing factors compete or complement each other in regulating splicing.
- Another SNP and PTEN mutation: large effect. The SNP affects genes in the mTOR pathway. Model: synergistic interactions between the gene and PTEN lead to high activation of mTOR signaling.
- Conditioned on genetic background: search for increased mutations in given genes and found some pairs. The analysis stratifies tumors by germline variants (28 SNPs), combining all tumor types.
- Lesson: how cancer develops is quite sensitive (large effect size of common SNPs) to genetic background. If so, wed expect significant epistasis between somatic drivers and somatic non-coding mutations as well (a non-coding mutation with similar effects on gene expression would predipose the same kind of driver mutations).

The inherited genomics of childhood leukemia: From biology to clinical implementation [Jun Wang from St. Jude, 2017]

- GWAS of ALL susceptibility. Six loci, with OR = 1.5 - 2.5, larger effects than adult cancer.
- 5-10% children ALL have germline driver mutations. TP53, ETV6 (from families). The role of ETV6 is confirmed by somatic mutations. Normal function of ETV6: required for hematopoiesis and maintenance of the developing vascular network.
- Sequencing study: 4000 pediatric ALL subjects, 40 ETV6 variants, 31 predicted to be ALL-related. Half of mutations are in ETV domain (DNA-binding). Confirm the function of 21 ALL variants: drive reporter expression.
- TP53 mutations: 2% of all patients have TP53 mutations. Deleterious mutations: CADD > 15, PPH2 and SIFT, and ExAC AF. Many TP53 pathogenic mutations are in DBD. TP53 pathogenic variants: associated with hypoploidy and with inferior survival.
- IKZF1: TF important for B cell development.
- Lesson: variations in genes affecting normal development can predispose to cancer. This could be different from adult tumor, where mutations target stem cells, or maintenance of cell states, etc.

Obesity gives unexpected boost to anticancer drugs [Science, 2018]

- Obesity patients are significantly more responsive to anti-PD1 therapy.
- Possible model: T cells are exhausted in obesity mice, and display more PD1 expression. So they are more easily suppressed by cancer cells. A similar model occurs for NK cells.
- Remark: possible mechanism is, obesity promotes inflammation, which suppresses CTLs.
- Remark: environment, genetics can both modify how tumor interacts with immune system. This is one example: obesity affects T cell states. Similar mechanisms may be found for germline variants.

Chapter 4

Cancer Functional Genomics

Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins [Rozenblatt-Rosen & Vida, Nature, 2012]

- Idea: DNA tumor virus (e.g. HPV) perturb the cellular networks to make it tumor. Two types of perturations: expression of viral genes and virus-host protein interactions. These perturbations should be similar to those from somatic mutations. The advantage of studying cell perturbation using virus is: (1) the targets of viral proteins can be determined experimentally and these targets are good candidates of "driver mutations"; (2) the viral gene can be introduced one at a time to study which one drives cancer. In contrast, in somatic mutation data, causal and driver mutations are not easily distinguished.
- PPI network between host and virus through Y2: viral-host interaction network of 454 binary interactions involving 53 viral proteins and 307 human proteins. Analysis identified 31 host target proteins that exhibited more binary interactions with viral proteins than would be expected given their "degree".
- PPI network between host and virus through mass spec.: introduce viral gene into host cells and study protein complexes by mass spec.
- Analysis of host targets: e.g. the transcriptional regulators CREBBP and EP300 were found to associate with E6 proteins from both cutaneous HPV types, but not with those from the mucosal classes.
- Host transcriptome perturbation: (1) Model-based clustering of the 3,000 most frequently perturbed host genes identified 31 clusters. (2) The mean expression change of each cluster revealed three distinct groups of viral proteins. Ex. Group III viral proteins increased expression of genes that are involved in cell proliferation and whose promoters are enriched in E2F binding sites.

miR-196b directly targets both HOXA9/MEIS1 oncogenes and FAS tumour suppressor in MLL-rearranged leukaemia [Li & Chen, Nature Comm, 2012]

- Problem: the role of miR-196b in AML, how it regulates other AML-related genes?
- miR-196b represses expression of HOXA9 and MEIS1 (known oncogenes of AML). Experiment: direct targeting of 3'UTR of the genes, and reduced expression.
- Observation: overexpression of miR196b is associated with adverse survival in AML patients. Why?
- FAS is associated with AML survival, and it is a direct target of miR196b. Mechanism: FAS promotes apoptosis.
- Model: miR-196b has dual role in hematogenesis. HOXA9/MEIS1 stimulate cell proliferation and FAS stimuates cell differentiation. miR-196b regulates both steps. Perhaps this allow the body to control the balance between two processes with a single molecule.

- **Lesson**: cancer is related to the dysregulation of the control of cell development/differentiation process.

High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities Screens Reveal Fitness Genes [Hart & Moffat, Cell, 2015]

- Comparison with RNAi: off-target effects, incomplete knockdown.
- How genetic screens can be used? Fitness genes, synthetic lethality, drug resistance, metastasis, immune response.
- Data: 5 cell lines, about 6 gRNAs per gene (12 in a larger library). Total of 4,000 hits, 4-5 times more than RNAi screens. Likely explanation: RNAi only good for highly expressed genes.
- BAGEL algorithm: obtain $P(D|\text{ess})$, where D is read count data $P(D|\text{non-ess})$ and compute BF. The distributions can be trained using gold-standard genes.
- Core fitness genes: about 1,500 in 3/5 cell lines. Mostly represent core processes such as RNA splicing, translation, DNA repair (4-fold enrichment). Also depleted of cell-cell communication and developmental process.
- Studying uncharacterized core fitness genes: confirm the effect by knockout, then subcellular localization and PPI (IP followed by Mass Spec). Show interactions with RNA splicing, protein folding.
- Context-specific fitness genes: e.g. telomere maintanence only in TERT-transformed cell lines. ETC complex in one cell line. Even in two similar cell lines, could have different fitness genes (e.g. RTK genes).

4.1 Cancer Epigenomics

Epigenetics: Separate paths for epigenomes and genomes in cancer evolution? [NRG, 2016]

- AML: genetic mutation load is low comparing with solid tumors, and the mutations often affect epigenetic regulators.
- Epigenetic heterogeneity (assessed using BiS-seq) in AML: increased, and associated with poor survival.
- Role of genetics: Increases or decreases in epigenomic diversity between diagnosis and relapse were not accompanied by equivalent changes in genetic diversity.

The chromatin accessibility landscape of primary human cancers [Corces and Chang, Science, 2018]

- Data: 400 samples (primary), ATAC-seq, from 23 tumor types. Average about 100k peaks per tumor type.
- Comparison with OCRs in normal tissues: About 66% overlap with Roadmap data, and the rest is new. Overlap of specific tumor types vs. normal tissue types: generally highest among matching tumor/tissue types.
- Correlation of chromatin profiles and sample properties: (1) Myc example, all OCRs in the locus, clearly two distinct clusters across 23 tumor types. (2) Tumor type specificity of OCRs: distal OCRs more specific than promoter OCRs, or gene expression. (3) Cancer subtypes: in some cases, see several subtypes of cancer based on OCR profiles.
- Clustering analysis of OCRs: performed only on cancer type specific OCRs (via clustering analysis) - focusing on distal OCRs (>100kb away). Assign clusters of OCRs to tissues using Roadmap data.
- Molecular mechanisms of chr. accessibility: Motif analysis. Clusters of OCRs (across all samples), and motif enrichment analysis. Also correlation of [TF] with OCR cluster activity across samples.

- Molecular mechanisms of chr. accessibility: TF footprint analysis (Figure 4AB). The hypothesis: TF variation drives chr. accessibility. Two measures of footprints: (1) flanking accessibility: $\log_2$ flanking / background; and (2) protection: $\log_2$ footprint / flanking. For each TF, correlation of [TF] with average flanking accessibility; and with protection. The former can tell if a TF opens or represses chromatin. Examples of repressive TF: CUX1.

- Molecular effects of OCR: target genes, by correlation of ATAC peaks vs. gene expression (500kb). Found average 5 OCRs per gene, or 1.6 genes per OCR. Only 24% are nearest genes.

- Identifying OCRs related to immunological status: correlation of peaks vs. cytotoxic activity across samples. PD-L1 OCR show significant correlations.

- Putative noncoding elements in cancer from somatic mutations: (1) Define AS Chromatin variants: from WGS data, call somatic mutations. Then do AS analysis of all mutations in ATAC-seq data (Figure 7B). (2) Validation of ASC variants: correlation of variant status vs. gene expression across samples, TERT, FGF4D (Figure 7CD). (3) Disruption of motifs in the examples.
  Note: need WGS as control because of copy number changes.

- Lesson: when integrating data of different sources, one often needs to match tissue/cell types. This may not always be straightforward. One can do this in a data-driven fashion: choose the tissue types that maximize some measure.

- Lesson: TF footprint analysis, (1) two separate metrics, flanking accessibility and protection. (2) Correlation of [TF] and TF footprint accessibility: can confirm the functions of TFs in controlling chromatin accessibility and the effect direction.

- Lesson: Validation of ASoC variants using eQTL across samples.

Chapter 5

Tumor Heterogeneity and Evolution

Cancer evolution [person notes]: the main challenges.

- How often does an adaptive mutation (driver) emerge? Can a driver mutation have large fitness effect or cause a large shift of cellular behavior?
- What is the strength of selection relative to population size and time? Does selection often drive to fixation?
- How important is epistasis, or how evolutionary trajectory is contingent upon the past events?
- Does similar selective pressure drive the same genetic changes, aka. adaptive convergence?
- Implications on the pattern of genetic divergence of cells within a tumor? Are they often the results of a single clone, driven by a few adaptive mutations?
- A test of selection based on divergence pattern within cells of a tumor? Benefit: discovery of patient-specific driver genes/events. Idea: Parallel evolution in multiple subclones may suggest positive selection.

Tumor heterogeneity [Personal notes]

- Why heterogeneity? (1) Relatively small time, comparing with population size. (2) Growing populations.
- Representation of the tumor evolution process: mutation tree, each node represents a mutation, analogous to speciation (creating a new clone). Representing time and cell numbers of each clone.
- Key challenge: inferring fitness of mutations. Neutrality assumption for most mutations.
- Advances in understanding tumor evolution through single-cell sequencing [Kuipes, BBA, 2017]
- Difficulty of reconstructing tree from bulk data: identfiabilty problem. Ex. suppose $f_a > f_b$ and $f_a > f_c$, then we can order the mutations a vs. b and a vs. c, but we cannot determine the order of b and c.
- Sum rule: the total AF of a node must be equal to or greater than the AFs of all child nodes.

The impact of recombination on selection/analysis of KaKs ratio [personal notes]

- Mullers ratchet: in asexual organisms, under the assumption of finite population and high mutation rates of deleterious mutations, the genetic load of the population will grow and eventually lead to extinction.

- Clonal interference or Hill-Robertson effect: in asexual populations, beneficial mutations/clones can compete with each other. In contrast, recombination facilitates the combination of beneficial alleles, making them faster to fix.

- Question: in asexual populations with both positive and negative selections, what are the rates of fixation? Will the rate of fixation of positive selection be reduced by deleterious mutations?

- Analysis: suppose the rate of deleterious mutations is high. In clones with a beneficial mutation, at some point, each cell will contains some deleterious mutation, which effectively cancels out the fitness advantage of the beneficial mutation. As a result, the rate of fixation of positive selection can be delayed.

- Issue of population size: the population size is small initially, but can become very large over time. This influences the model assumptions.

Only three driver gene mutations are required for the development of lung and colorectal cancers [Tomasetti and Vogelstein, PNAS, 2015]

- Model of incidence time: suppose cancer initiation requires n driver mutations in a certain order. Let u_i be the rate of i-th driver event, then the waiting time for the i-th event X_i follows exponential distribution with rate u_i, so the total waiting time is $I = \sum_{i=1}^{n} X_n$. Using the PDF of the sum of exponential RVs (note: sum of expo. distribution with identical rate follows Gamma distribution), we have the PDF of the waiting time follows:

$$I(t) = u_1 \cdots u_n \frac{t^{n-1}}{(n-1)!} \tag{5.1}$$

  It is easy to check that when $u_1 = \cdots = u_n = u$, this reduces to Gamma distribution Gamma(n, u). The waiting time distribution translates directly to incidence rate: at the log-scale

$$\log I(t) = \log \frac{u_1 \cdots u_n}{(n-1)!} + (n-1) \log t \tag{5.2}$$

  So the slope of the log-incidence rate vs. log-t gives the number of required driver events.

- Remark: an alternative model would ignore the order of events. Suppose we have n driver mutations, and any of m mutations would be enough to initiate tumor. We can model the rate of driven mutation as: $\lambda n/N$, where λ is per base mutation rate, and N total number of possible mutations. Then the number of driver mutations in time t follows Pois$(\lambda t n/N)$.

- Extension of basic models: different types of mutations, variable number of driver mutations. Ex. with n different driver genes of different length $l_1, \cdots, l_n$, and each base has equal mutation rate u, then we have $u_i = l_i u$. Ex. suppose n can vary from m to M, and there are different combinations of n mutations that lead to cancer. Let j_n be one combination of n drivers, we have:

$$I(t) = \sum_{n} \sum_{j_n} l_1(j_n) \cdots l_n(j_n) u^n \frac{t^{n-1}}{(n-1)!} \tag{5.3}$$

- Effect of mutation rates: suppose we have two subgroups, in one group, mutation rate is x times higher than the other for each driver event. Then the incidence rate of the group is x^n times higher than the other group. Advantage of focusing on comparison of incidence between groups: robust to violations of model assumptions.

- Dependency of mutations: each mutation will increase fitness. Let λ_1 be the growth rate of first driver, λ_2 be the second, and so on, with $\lambda_1 < \lambda_2 < \cdots < \lambda_n$. Effectively, we can think of this as: positive

selection increases effective mutation rate in each step (or reduce waiting time). So the mutation rate factor in the incidence time equation becomes:

$$u_1 u_2^{\lambda_1/\lambda_1} u_3^{\lambda_1/\lambda_2} \cdots u_n^{\lambda_1/\lambda_{n-1}} \tag{5.4}$$

Under a simple model that each driver increases equally the fitness, we have: $\lambda_1/\lambda_2 = 1/2$, $\lambda_1/\lambda_3 = 1/3$, and so on. This implies that if mutation rate is x times higher in one group, the incidence rate would be:

$$x \cdot x^{\lambda_1/\lambda_1} x^{\lambda_1/\lambda_2} \cdots x^{\lambda_1/\lambda_{n-1}} = x \cdot x \cdot x^{1/2} \cdot x^{1/(n-1)} \tag{5.5}$$

Note: this makes sense, since the rate increase would be smaller than x^n.

The Ecology and Evolution of Cancer: the ultra-microevolutionary process [Wu, ARG, 2016]

- Main difference between cancer evolution and organism evolution: (1) very short divergence. (2) Evolution is massively reiterated, and could answer the question of adaptive convergence. (3) Mutation rate itself in cancer is evolving: needs to be considered, e.g. high mutation rate can lead to mutational meltdown and population extinction.

- Stage I evolution: tumor vs. normal cells (similar to inter-species comparison). Stage II evolution: within tumor cells (similar to intra-species comparison).

- Stage I: low genetic convergence. Modest overlap in genes among cases of the same type, and also modest overlap in genes of different tumor types. Ex. APC 92% in colorectal cancer vs. 7% in others; VHL 52% in kidney vs. 7% in others.

- Analysis of adaptive evolution in cancer in Stage I: the enrichment in cancer vs. background, λ, is simply K_a/K_s ratio.

- Population genetic analysis: consider only fixation, the probably of fixation for an adaptive mutation is $2s$ and for neutral is $1/N_e$. For a gene, suppose in non-syn. mutations, fraction p, q are positive and negative selection, its easy to show that $K_a/K_s = (1 - p - q) + p2N_e s$. If we take average s over all mutations of a gene, $K_a/K_s = 2N_e s$.

- Empirical pattern of K_a/K_s: Average K_a/K_s is 1 for cancer vs. 0.21 for human-primate comparison. This suggests that (1) Negative selection is likely reduced vs. inter-species comparison; and (2) positive selection overall balanced by negative selection. Compare the distribution vs. neutral expectation: enrichment of both positively selected (5%) and negatively selected, $K_a/K_s < 0.5$ genes (about 15% or more).

- Discussion of weak selection in tumor: small N_e, especially for solid tumors.

- Stage II: evidence supporting neutrality of intra-tumor diversity. Very high diversity: Ling et al, estimates 100M mutations. Comparison of the pattern vs. expected under neutral model (infinite-site or infinite-allele model): consistent with neutrality - no unusually large clones. However, power consideration is important. Williams et al uses the mutations that are locally polymorphic but globally rare for testing neutrality.

- Remark: Interpretation of K_a/K_s pattern: use average value for genes. The argument of weak selection may be questionable: many positions are highly selected (recurrent mutations) - strong evidence of adaptive convergence. So for some positions, K_a/K_s could be very large, and this would drive dominance of a single clone.

- Remark: how to explain K_a/K_s is close to 1? How much is explained by positive selection vs. greatly reduced negative selection?

- Remark: lack of recombination makes it difficult to test positive selection in data from an individual sample, even if we have single-cell sequencing data, because of genetic hitchhiking.

Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine [Lipinski and Gerlinger, TIC, 2016]

- Problem: prediction of cancer response to treatment and whether it will recur.
- Mutational process: 20 mutational signatures found, 9 of which can be linked to mechanisms, e.g. APOBEC introduce hypermutation clusters. Chromosome abnormality: e.g. DNA fragments detached (EGFR). Mutational process can differ between patients and stages. Interaction between mutation and driver mutations: mutation mechanism may predispose certain kind of mutations, leading to recurrent driver mutations, e.g. APOBEC mutations favor PI3K mutations.
- Random drift and population size: what matters is the number of cancer stem cells. Population bottleneck: e.g. after treatment.
- Selection: competition between subclones. Spatial constraint limits the competition, s.t. a highly beneficial mutations may not reach 100% frequency. Selection factors in TME: e.g. blood vessels, immune cell infiltrations.
- The emergence of drug-resistant clones: EGFR inhibitor, the resistant clones often have T790M mutations. Another example, resistant clone of EGFR with KRAS mutations exist before treatment.
- Cancer from population genetics view I: the impact of mutation rates and population size. Some cancers, e.g. CML: low mutation rates (genome stable) and small population size, random drift important. Often in solid tumors, late stages, almost any mutations will occur.
- II. Interaction of selection and mutation rates/population size. Cancer genome doubling or chr. duplication may reduce selective constraint. The role of mutation rates and population size (Figure 3): increase mutation, increase mutation load as well as chance of beneficial mutations, so larger variance of fitness; increasing population size increase mutation supply and increase combination of beneficial mutations (also increased by catastrophe events).
- III. Mullers ratchet: could be relevant, accumulation of deleterious mutations. Evidence that hypermutable tumors show better prognosis.
- IV. Fitness landscape: need to reconstruct fitness landscape of a tumor using data from many patients/tumors. Important part is epistasis, including synthetical lethality.
- Strategies of improving predictability: macroheterogeneity profiling (multi-region), microheterogeneity profiling (single-cell), ctDNA to monitor dynamics.
- **Lesson**: understand genetic heterogeneity of cancer using a population genetic perspective. (1) Tumors often use different strategies to evolve a hallmark: this could be the result of random drift during early progression when effective population size is small. (2) Selection may not be effective because of spatial limitation. (3) In late stage with treatment, large selection, large population size, then results may be more deterministic.

Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future [McGranahan and Swanton, Cell, 2017]

- Part I. Pattern of ITH: high burden of clonal mutations (e.g. lung, melanoma) may have lower ITH. Clonal mutation burdens reflect mutations accumulated prior to tumorigenesis.
- Driver mutations in ITH: subclonal driver mutations can give an illusion of colonality due to sampling bias. Some driver mutations tend to subclonal, e.g. PI3K-AKT-mTOR pathway, comparing with RAS-MAPK pathway.

- Part II. Tumor evolution. Neutral model or selection: (1) Evidence of neutral model: the relationship between the number of subclonal mutations and their relative abundance. (2) Evidence of selection: driver genes tend to harbor more subclonal mutations than expected.
- Contingency: e.g. in myeloid cancer, the order of JAK2 and TET2 matters - different clinical phenotypes.
- Convergence: parallel evolution of multiple mutations targeting the same genes in different subclones. Many examples, e.g. VHL cancer, multiple SETD2 mutations in several branches of the tree.
- Punctuated evolution: macroevolution events such as genome doubling and chromothripis (cluster of rearrangements).
- Metastasis as speciation: many possible scenariors, one subclone from primary sites creating metastasis in multiple sites; or multiple subclones from primary sites in different metastatic sites; or polyclonal metastatsis.
- Part III. Tumor ecosystem. Functional cooperativity between subclones: some minor subclones promote the survival of main subclones, e.g. EGFR-mutant subclone supporting EGFR-wt main clone via paracrine stimulation (LIF, IL6).
- TME: e.g. hypoxia condition can influence tumor survival. Safe havens created by tumor cells: e.g. tumor cell may promote remodeling of extracellular matrix.
- Part IV. Treatment taking into account ITH: targeting driver mutations or neo-antigens in the dominant clones.
- Adaptive therapy that deals with competitive release of resistant subclones: when we remove the major subclone, the resistant subclone is promoted because of lack of competition.
- Targeting tumor genome instability: e.g. PARP inhibitors in BRCA patients.
- Exploiting evolutionary constraint: knowledge about parallel evolution/convergence of tumor.
- Q: Why tumors with high clonal mutation burdens, including lung cancer and melanoma, have lower ITH?
- Q: Detection of neutral vs. selection: comparison of subclonal mutations and their frequencies can suggest subclonal expansion. However, it could also be explained by earlier mutational events in some subclones?
- Remark: convergence is a key evolutionary property of tumor. Can we detect convergence at a higher level than genes? Ex. epigenomic/transcriptomic alternations; intermediate traits.

Quasi-neutral molecular evolution When positive and negative selection cancel out [Chen and Chung-I Wu, NSR, 2018]

- Relationship of dN/dS ratio and selection: let $R = dN/dS$, p, q are fractions of positive and negative selection, we have $R = 1 + p * (2Ns - 1) - q$, where $2Ns$ is strength of positive selection. For most tumors, R is close to 1.
- The existence of negative selection: use $2Ns > 3$ (small value), and $p = 0.05$, we have $q = 0.1$. More generally, to explain $R = 1$, we need $q = p(2Ns - 1)$, so q is generally significantly higher than p.
- Influence of stem cell population size: comparison of small intestine vs. colon, population size 2 - 50. Claim: difference of population size explains the efficiency of selection and difference of cancer rate.

- Q: Population sizes can be very different, then $2Ns$ can be very different, but R is roughly similar across cancer types. Why?
- Q: If negative selection is common, why we havent found much? Possible answer: spread in many genes, and within a gene, the difference is not large.
- Remark: population size is only based on early stage evolution.

Statistical problems in cancer genomics [Simon Tavare, 2019]

- Big Bang model: most clones occur very early. Neutral evolution model: Williams, 2016. Not much selective sweep.
- Coalescence of cancer: at cell level. Infinite site model.
- Cancer sequencing data: mutation x cell matrix. Summary stats: haplotype freq. spectrum. SFS.
- Simulation of SFS: sample mutations in the branches of the tree.
- Dahmer and Kersting (2015): distribution of SFS, converges to normal aymptotically. Each bin (of frequency) would behave like independent Poisson. Note early methods (Fu 2015) only shows expectation.
- Binned SFS: 20K SNVs. Study one chromosome. Can we estimate number of sweeps from the data?
- Problem: infer distribution of K (number of haplotypes) and theta (mutation rate) from SFS.
- ABC: simulation data with theta, and compare with obs. If close, keep theta.
- To use ABC for our problem: match S (number of segregating sites) and SFS. Ex. difference of S vs. S.obs; or KS test of SFS.
- Model with selection. Show that SFS would look different.

5.1 Tumor Heterogeneity and Phylogeny Methods

Principles of reconstructing the subclonal architecture of cancers [Dentro and Van Loo, CSHL Perspective, 2017]

- Concepts: VAF, CCF (cancer cell fraction), purity, multiplicity. Note: VAF, CCF and multiplicity are defined for mutations, while purity is for the entire sample.
- Relationship between VAF and CCF, purity and multiplicity: let ρ be purity, π_i be CCF of mutation i, m_i be its multiplicity in tumor cells with the mutation (e.g. normally 1, if duplication after mutation 2, 3, etc.). Let t_i and n_i be number of chr. copies of tumor and normal samples, respectively. Let f_i be the VAF, then
$$\mathrm{E}\left(f_i\right) = \frac{\rho\pi_i m_i}{\rho t_i + (1-\rho)n_i} \tag{5.6}$$
where the numerator is the number of chr. copies with the mutations and the denominator is the total number of chr. copies. Note: under infinite site model, each mutation occurs only once, so m_i has a single value. CNVs could complicate the problem.
- SNV-based subclonal reconstruction: model read count as binomial with probability of alternative alleles $p_i = \xi_i\pi_i$ where ξ_i is the fraction if mutations are clonal and π_i the CCF (above). Clustering of mutations are based on CCF, so use Dirichlet Process prior for π_i.

- Copy number analysis: using BAFs to call CNVs. Assuming the CNVs are clonal, the BAF of a CNV similarly depends on purity and copy number in tumor cells t_A, t_B:

$$BAF = \frac{\rho t_B + (1-\rho) n_B}{\rho(t_A + t_B) + (1-\rho)(n_A + n_B)} \tag{5.7}$$

  where n_A, n_B are copy number in normal cells. BAF plot (Figure 4): the shape (number of bands) reflects copy number changes, and the values of BAF reflect tumor purity and CCF. The bands become blurred at lower purity and subclonal CNVs at lower coverage (40x vs. 100x). Refining CNV analysis by using haplotype phasing.

- Inferring subclonal CNVs: possible from BAF analysis. However, difficult, because number of subclones and the copy number of a CNV are coupled. Some assumption may help resolve ambiguity: e.g. only two subclones.

- Phylogeny reconstruction: grouping mutations by CCF. Mutation phasing of SNVs and CNVs help: however, generally rare.

- Remark: for detecting subclonal CNVs, WGS has large advantages over WES (large BAF signal).

Computational approaches for inferring tumor evolution from single-cell genomic data [Zafar and Chen, COSB, 2018]

- Experiment for single-cell-sequencing (SCS): different options for WGA, e.g. in vitro transcription leads to linear amplification.

- Errors in SCS: (1) non-uniform coverage: creating problem for CNA. (2) Doublet cells: more than 1 cell in a well, about 1-10%. (3) Allelic dropout (ADO), locus drop (not amplified), allele imbalance and sequencing error.

- Variant detection: (1) CNAs: most method rely on coverage uniformity. (2) Monovar: multi-sample to overcome coverage nonuniformity. (3) SCcaller: single sample calling. (4) Directions: combine bulk and SCS data, using phylogeny.

- Reconstruction of subclones: dimensionality reduction and clustering methods to explore mono- vs. polyclonality.

- Tumor phylogeny: (1) OncoNEM: relationship of subclones, marginalize mutation placement, greedy search. (2) SCITE: mutation tree. (3) SiFit: cell tree, finite-site model that account for mutation recurrence and loss. Mutation loss is probably important: it happens relatively common, and if not remove, will distort the tree.

The Life History of 21 Breast Cancers [Nik-Zainal and Campbell, Cell, 2012]

- Data: 1 cancer genome, sequenced at 188x. 20 genomes, sequenced at 30-40x.

- Copy number profile (Figure 1A): both logR and BAF, large CNVs show distinct signatures. Subclonal CNVs show somewhat lower logR, and lower BAF bands than expected.

- Expected BAF (personal notes): we consider only germline SNVs, so only copy number changes matter. Let ρ be tumor purity, ϕ be the percent of tumor cells carrying a copy number change, and C_A, C_B be the copy number of A and B alleles in the tumor cells with copy number changes, N_A, N_B be the copy number in other tumor cells and normal cells. The expected BAF is:

$$\text{BAF} = \frac{\rho\phi C_B + \rho(1-\phi) N_B + (1-\rho) N_B}{\rho\phi(C_A + C_B) + \rho(1-\phi)(N_A + N_B) + (1-\rho)(N_A + N_B)} \tag{5.8}$$

  Note that BAF is not a linear function of ϕ. Ex. if subclonal chr. deletion of B alleles in 40% cancer cells, we have: $\phi = 0.4$, $C_A = 1, C_B = 0$. Or subclonal chr. gain of A alleles in 30% of cells, we have: $\phi = 0.3$, $C_A = 2, C_B = 1$.

- Haplotype based CNV analysis (Figure 2A): Battenberg patterns of AFs of variants in the same chromosome. Within a haplotype block where phasing is done, we see lines of one parent alleles; but across blocks, hard to phase. This leads to the Batternberg pattern. Allows one to estimate tumor cell fraction of subclonal CNVs.

- Clusters of SNVs by VAFs (Figure 1B): some small clusters of SNVs (by VAFs and read counts) are mutations in CNVs. Found 4 clusters.

- Resolving relations of SNVs and CNVs (Figure 2B): use somatic mutations and phased heterozygous SNVs to determine the order of SNVs and CNVs. Ex. subclonal deletion of chr. 13, analysis of SNV VAFs if they occur before the deletion; or after the deletion. The VAF of a SNV depends on which parental copy it occurs.

- Phasing somatic mutations to resolve ambiguity of phylogeny: even though a very small number of somatic mutations can be phased, they are informative for subclonal relationship. Ex. two VAF clusters, B and C with VAF 12% and 20%. If we never see any mutations in the B and C clusters together, it means that the two clusters are likely in separate branches.

- Phylogeny of 20 breast cancer: (1) Timing chr. evolution in 20 breast cancer genomes: before the first major CNV, some SNVs already occured, and CNVs continuous ongoing. (2) All samples have a dominant subclone ($> 50\%$), and for a few samples, can construct phylogeny.

PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors [Deshwar and Morris, GB, 2015]

- PhyloSub model: how DP prior is used in the context of a tree? Basically, the VAF in the root node is Uniform(0,1), and for any other node ν, the VAF satisfies the constraint that it must be lower than the VAF of the parent node.

- Background: infinite site assumption. Ambiguity (Figure 1D), rely on parsimony assumption: favor small number of populations, especially no vestigial VAF clusters (zero frequency). Note: the parsimony assumption is encoded by the prior of subpopulations.

- How copy number affects the VAF of SSMs (simple somatic mutations)? Figure 2.

- Incorporating CNVs with SSMs: (1) CNV input: C_i, ϕ_i, the copy number and population frequency. (2) Treat CNVs as pseudo-SSMs. (3) Adjust for population frequency of SSMs that overlap with CNVs. Three scenarios.

- Simulation study. Results: WGS of 30-40x can resolve 3-4 subpopulations. Limit to SSMs in normal copy number regions leads to reasonably good results.

- Benchmarking: TCGA benchmark for variant calling, mixture of several populations.

- Breast cancer data: among 26K SSMs, 4.7K are in clonal CNVs, and 2.1K are in subclonal CNVs. Or 26% of genomes are affected by CNVs, and about 1/3 of CNVs are subclonal.

Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing (Canopy) [Jiang and Zhang, PNAS, 2016]

- Input data of Canopy: (1) SNV: read count matrix R (for each SNV in each sample), and coverage X. (2) CNV: major copy numbers W^M and minor copy numbers W^m. These are continuous values (average copy numbers in the sample) from allele-specific copy number estimation programs.

- Model structure: let τ_K be the tree with K subclones and P the fraction of each subclone. The leftmost branch is the normal cells. Let Z be the genotype matrix of SNVs (SNV presence in each subclone), and $\tilde{C}^M, \tilde{C}^m$ the genotype matrices of CNVs (CNV presence in subclone). Our model for SNVs:

$$\tau_K, P \to Z, \tilde{C}^M, \tilde{C}^m \xrightarrow{\tilde{H},\tilde{Q}} VAF \xrightarrow{X} R \tag{5.9}$$

  where $\tilde{H}$ denotes temporal phasing (order of events) and $\tilde{Q}$ whether SNVs and CNVs overlap. Note that VAF depends on the ordering of SNV and CNV events. For CNVs, our model:

$$\tilde{C}^M, \tilde{C}^m \to C^M, C^m \to W^M, W^m \tag{5.10}$$

  where C^M, C^m are the expected major and minor copy numbers. In the model, the genotype matrices of SNVs and CNVs are entirely determined by the tree, subclone percent.

- VAF model: if SNV is not in a CNV, then VAF is simply the frequency of subclone divided by 2. If it overlaps, there are three cases: before CNV, after CNV, or two independent events (in separate branches).

- Inference: MCMC, infer number of subclones K, their frequencies P, SNV and CNV placement in the branches (genotype matrices).

- Simulation: (1) under a fixed history (Figure 1), sample coverage from multinomial distribution, and analytically calculate VAFs from the equations. Then sample read counts from binomial. (2) Randomly place mutations: a fixed tree topology and subclones.

- Simulation results: evaluation by errors in genotype matrix Z, and RMSE of P. Figure 3, and Table S3. Generally don't need many mutations, in Figure 1 tree with WGS $d = 30$, with < 15 mutations can achieve errors of Z less than 0.1.

Clonal genotype and population structure inference from single-cell tumor sequencing (SCG: Single Cell Genotyper) [Roth and Shah, NM, 2016]

- Model idea: cluster single cells into a finite set of populations, assuming each population has the same genotype, and infer the genotype of each population. Phylogenetic interference can be done on the inferred populations (separate, downstream analysis).

- Model: each cell belongs to one population, and G_{km} be the genotype of m-th locus of population k. The observed genotype depends on the true ones. Use Dirichlet prior for the mixture weights.

- Multi-sample sequencing data from the same patient. Use topic model: each sample is a mixture of the same clones, with different weights between samples.

- Other extensions: doublet cells, multiple events (SNVs and breakpoint events), and position-specific errors (depending on ploidy).

- Results: 6 clones in single-cell WES data, and place driver events TP53 and ERBB2 (SNV and SV). While TP53 mutation is present in all cells, ERRB2 is only present in some clones (metastatic ones).

Tree inference for single-cell DNA data (SCITE) [Jahn and Beerenwinkel, GB, 2016]

- Mutation matrix (input data): possible errors include: (1) False positives: sequencing errors. (2) False negative: Allelic dropout. Also it may have missing data. Mutations present in all cells or in a single cell are not informative, and can be removed.

- Sample tree (cell tree): issue with identifiability. Ex. cells with identical mutations.

- Mutation tree representation. Each node is a mutation, and the tree shows the order/relation of mutations. Mutations shared by the exact same cells are not informative and can be combined into a single node. Also cell attachment to the nodes in the tree (could be attached to an internal node or even root). Note that: given a tree, the mutations of an attached cell is uniquely determined (the path from the root to the leaf).
- Subclone structure from mutation trees: group all cells with the same mutational path.
- Model: our observed mutation matrix is D. Let T be the tree, and σ be the cell attachment. The two determines the underlying mutation matrix E. The model of $P(D|E)$ is given by the error model (FP and FN rates, denoted as θ). We need $P(D|T, \sigma, \theta)$.
- Modeling homozygotous mutations: cannot happen under the infinite-site model, so treat them as errors.
- Inference: σ need not be explicitly sampled - it can be margalized.
- Application: 58 tumor cells, choose 18 cancer-related mutations. Estimate FP rate = 6E-6 and FN rate = 0.43, with 45% missing data. Most trees have linear chain of mutations. Increase to all 78 NS mutations, the posterior of tree is quite flat.

OncoNEM: inferring tumor evolution from single-cell sequencing data [Ross and Markowetz, GB, 2016]

- Model idea: multiple clones (subpopulation of cells), and mutations originate from clones. Because mutations do not occur in edges, we need to account for that by having unobserved clones. Ex. suppose we have two clones with mutations AB and AC, we need to have a clone with mutation A.
- Preprocessing: constructing genotype matrix, using consensus filter, i.e. a mutation is called only when it's found in more than 2 cells.
- Model: our data is d_{kl}, the genotype matrix of cell k in position l. We assume there are N clones. Let δ_{kl} be the true genotype. Our main parameters are $\theta_l \in \{1, 2, \cdots, N\}$, the clones where a mutation originate. Given θ, the true genotypes δ_{kl} are completely given. The genotype matrix d_{kl} is related to δ_{kl} by FP and FN rates. This allows us to specify the likelihood, $P(D|T, \theta)$. Since we are interested in only T, we'll marginalize θ, assume the uniform prior distribution, i.e. a mutation occurs in each clone with equal probability.
- Inference: first build a cell-tree by heuristic search - score the trees by likelihood; then merge nodes along a branch by evaluating the likelihood (if the likelihood ratio/BF is greater than some constant).
- FP and FN rates: note that they are different from FDR, and allelic dropout (ADO) rates. ADO is defined on any alleles, while FN rates are defined on the given genotype matrix (already filtered).
- Application to real data: WES data of 50 cells. Estimated FP and FN rates are aboud 0.2 and 0.08. In one sample, 3 subpopulations of cells.
- Impact of LOH: LOH can lead to loss of mutations, which is not allowed under infinite-site model. However, LOH can be tolerated as FNs (mutations not detected). In real data, removing the LOD region does not change much the tree.

SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models [Zafar and Chen, GB, 2017]

- Background: Perfect phylogeny: two cells are either related (subset of each other) or unrelated/disjoint. Under the infinite site model, for any two loci, we cannot have all four possible genotypes present - four gamete test (two mutations are either unrelated, or one occurs after the other).

- Model: let D_{ij} be observed genotypes for cell i and locus j, and G_{ij} be true genotypes. The model of $P(D_{ij}|G_{ij})$ is given by Equation (4). In particular:
  - $P(D = 1|G = 0) = \alpha$: FP rate, about 0.01.
  - $P(D = 0|G = 1) = P(D = 2|G = 1) = \beta/2$: FN rate, 0.2 - 0.4.
  - $P(D = 2|G = 0) = \alpha\beta/2$: ADO rate is $\beta/2$, and FP error rate α.

  For the evolution model, use continuous-time MC: the transition probability depends on rates and branch length. Fix the rate of $0 \to 1$, and define LOH and deletion rates, so $1 \to 0$ and $1 \to 2$ have rates $\lambda/2$. Likelihood is computed by Felsenstein's peeling algorithm, with the ancestral genotype fixed at 0.

- Convergent evolution: it is possible that the same locus mutates twice in two cells. Also back-substitution (1 to 0) is modeled by LOH and deletions.

- Inference: find the best tree and parameters using MCMC. (1) Tree: branch length change (scale all branches by a random variable); pruning-regrafting move, swapping move. (2) Parameters θ (error rates) and λ (evolutionary rates): Beta prior, and normal proposal distribution.

- Evaluating tree: RF distance, average of FP and FN rates. The rate is defined as: if an edge is present in the true tree, whether it is also present in the false tree (by comparing bi-partitions of leaf nodes of two trees - which has 1-1 relationship with edges).

- Simulation: give the root genotype, then place mutations along the tree. Allow back-mutations, deletions and LOH with certain probabilities. Also simulate doublets.

- Real cancer data: single-cell WES, 35 tumor cells. First show that 1800 pairs (out of 3000) of loci show evidence of finite-site (however, this could be due to sequencing errors).

- Remark: depending on which positions are included in the analysis, branch length has different interpretations. When all positions are included, branch length should be proportional to number of generations (cell divisions). However, when only variant sites are included, the branch length is a relative measure of the frequency of mutations in a branch.

- Remark: the method includes both LOH and deletion rates, but LOH is often caused by deletion (or UPD), so should not treat them as independent.

- Remark: the paper does not offer strong evidence that it's important to have finite-site model. The rate parameters are not reported, and the empirical four-gamete test does not account for sequencing errors.

SCIϕ: Single-cell mutation identification via phylogenetic inference [Singer and Beerenwinkel, 2018]

- Model overview: joint inference of genotypes and phylogeny. The read count depends on the underlying genotype via a Beta-Binomial model. Use cell tree, and marginalizing the mutation attachment.

- Read count model: suppose all candidate mutated sites have been identified (similar to previous work). Let c_{ij} be the coverage of locus i in cell j, and s_{ij} be the count of alternative allele. Let f be the underlying frequency, accounting for sequencing errors. We have $P(s_{ij}|c_{ij}, f_{ij}, \omega_{ij})$ follows Beta-Binomial distribution with ω overdispersion parameter. Use two different values for ω_{ij} depending on whether it is homozygous or mutant. The allele dropout (ADO) is modeled by: probability μ of ADO, mixture model.

- Tree inference: let D_{ij} be the data (both coverage and counts) and T cell tree. Let σ_i be mutation placement of site i, and θ be parameters. Let n be number of mutated loci and N be all sites. The model:

$$P(D|T,\theta) = \prod_{i=1}^{n} P(D_i|T,\theta) \prod_{i=n+1}^{N} P_{wt}(D_i) \tag{5.11}$$

  where the second term is used to estimate sequencing error rates. The first term:

$$P(D_i|T) = \sum_{\sigma_i} P(D_i|T,\sigma_i) \tag{5.12}$$

  where summation is done via Felsenstein's peeling algorithm.

- Incorporating LOH: assuming there is a certain fraction of LOH. Any site has a probability ν of being LOH. Mixture model with fixed proportion to account for LOH: if LOH, homozygotes; if not, the current model.

clonealign: statistical integration of independent single-cell RNA & DNA-seq from human cancers [Campbell and Shah, biorxiv, 2018]

- Method overview: from low coverage single cell sequencing, create single cell phylogeny, then obtain clones/clades (truncate). Do single-cell RNA-seq, and map to the clones. Given data Y_{ng} read count of gene g in cell n, and λ_{gc} the copy number of gene g in clone c, and the goal is to infer z_n, the clonal assignment of cell n.

- Model: expression of a gene depends on its copy number, let it be $\mathrm{E}\left(y_{ng}|z_n = c\right) \propto \mu_g f(\lambda_{gc})$ where $f(\lambda_{gc})$ is a function of copy number. The paper uses when $\lambda \geq 4$, $f(\lambda) = 4$, and otherwise, $f(\lambda) = \lambda$. To model expression read count:

$$\mathrm{E}\left(y_{ng}|z_n = c\right) \propto s_n \mu_g f(\lambda_{gc}) e^{\psi_n w_g} \tag{5.13}$$

  where s_n is a size factor for cell, ψ_n: random effect of cell n, and w_g random effect for gene g. The normalization constant sums over all genes. The distribution of y_{gc} is modeled with Negative Binomial, with overdispersion ϕ_{gc}. Use a hierarchical prior: $\log \phi_{gc} \sim N(\eta_g, \sigma^2)$.

- Inference: EM to sum over clonal assignment. Use Adam optimizer: high-dim. optimizer for DNN.

- Simulation: two clones, $N = 500, 1000$ cells, $G = 500$ genes. Vary the proportion of genes whose expression depend on copy numbers. Show that when the proprtion is higher than 0.5, very good performance.

- Validation: use a breast cancer dataset with 3 subclones. First cross-validation: hold out two chromosomes, use the rest to infer clones; then confirm that gene expression correlates with predicted copy numbers based on subclones. Next use LOH event on held-out chromosomes.

- Expression analysis: comparison of gene expression between subclones. Found that one major subclone: down-regulation of MHC class I and B2M (no copy number changes).

- Remark: the accuracy does not depend on CNV size or boundaries. What matters is the total number of genes with copy number changes in a cell.

5.2 Studing Tumor Heterogeneity and Evolution by Single-Cell Technologies

Unravelling biology and shifting paradigms in cancer with single-cell sequencing [Baslan and Hicks, NRC, 2017]

- Experimental design issues: sample quantity, e.g. may be difficult to get enough materials (for bulk) for pancreatic cancer. Tumor purity: significantly reduce the power of calling mutations.
- Cell-of-origin: subpopulations that give rise to tumor. Known cases: HSC and LDR5 stem cells in leukemia and intestinal cancer. ScRNA-seq study in normal sample can help define subpopluations.
- Pre-malignant tumor: limited by sample amount, and single-cell methods important.
- Primary tumor: Ex. pancreatic cancer: significant stromal infiltration poses challenge for bulk methods. Applications: (1) ScDNA-seq; Define key genetic alterations. (2) ScRNA-seq: define immunogenic subtypes of cancer.
- Metastasis: Origin of metastatic tumors: genetic adaptations or stochastic. Dynamic process: self-seeding. Stromal-tumor interactions: e.g. in prostate cancer, stromal cells support tumor growth.
- Implication of single-cell technologies on modeling cancer: organoid and CTC/DTC (disseminated tumor cells).
- Lesson: application of single cell technologies in (1) defining tumor heterogeneity: subpopulations. (2) Tumor microenvironment: immune subtypes, role of stromal cells in tumor. Both in primary tumor and metastasis.

Clonal evolution in breast cancer revealed by single nucleus genome sequencing [Wang and Navin, Nature, 2014]

- Method: use G2/M phase nuclei (more DNA). Use MDA: control the time to reduce FP rates.
- Method validation: about 50 single cells (WES or WGS or CNA), evaluation using a monoclonal cell line. Genome coverage breath: 91%. Replication across cells: $R2$ 0.9. ADO rate: 10%. FP rate due to sequencing is also low.
- Application to an ER+ tumor patient: CNA profile suggests two subclones. Most mutations found in bulk sequencing are found as clonal in single-cell data.
- Application to TN breast cancer: three subclones in both CNA and single-cell SNV data.
- Mutation rate estimation: use birth-death processes, and estimate mutation rates at single-cell level. TNBC 13 times higher than ER+ cancer.
- Lesson: evaluation of the quality of single-cell sequencing data by: genome coverage breath, reproducibility (across cells and vs. bulk), ADO rate, FP error rate.

Single-cell sequencing maps gene expression to mutational phylogenies in PDGF and EGFdriven gliomas [Muller and Diaz, MSB, 2016]; CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones [Diaz, Bioinfo, 2018]

- ITH in gliobastoma multiforme (GBM): amplification of RTKs are common, but they are often mosaic and regional. Treatment can induce ITH/clonal evolution.
- Characterizing three GBM patients with bulk WES and scRNA-seq: 288 cells. (1) Somatic mutations (Circos plot): EGFR, PDGFRA and KLHL9 mutations. (2) Tumor subtyping at individual cell level: by mutations, EGFR-driven or PDGF-driven; or by developmental origins, mesenchayl/classical or proneural/neural.
- Calling CNVs from scRNA-seq: (1) CNVs from bulk-WES: median size of 300-400 genes. (2) Use a reference (normal brain) scRNA-seq as control. For each cell, compare its normalized expression (log2 CPM) of every cell with the distribution in the reference, and use 5% in reference as cutoff.

- Phylogeny reconstruction using CNV genotype: a standard package, Phylip R. Example of independent deletion in chr 13 in two subclones.
- MiRNA cluster in chr 13 deletion: comparison of subclones with the mutation and without mutation, found DEX genes highly enriched with miRNA targets or targets of miRNA-regulated TFs. Also found DEX genes may increase cancer cell infiltration (using a reference cancer expression atlas).
- PDGFRA mutation: a in-frame deletion, very common. Show the dosage of the mutation correlate with expression of PDGFR pathway, cell cycle, etc.
- Transcriptome dynamics: follow the expression of the backbone of the tree (dominant clone). Group cells with the same CNV profiles into early, mid and late (progressive CNV gains). In one patient, constitutive expression of NSC genes, but gradual induction of OPC genes (also regulated to neuro-differentiation); also induction of PI3K/AKT pathway and angiogenesis.
- Lesson: characterizing cancer cells (sub-typing) by their expression signatures: mutation driven or developmental origin.
- Lesson: projected expression on mutation trees can be used for studying (1) Transcriptome dynamics: how expression of cancer-related genes evolve over time; (2) impact of specific mutations: correlating mutation status (or expression level of genes containing the mutation) with gene expression.

Integrated single-cell genetic and transcriptional analysis suggests novel drivers of chronic lymphocytic leukemia [Wang and Catherine Wu, GR, 2017]

- Background: intra-tumor heterogeneity at multiple levels (DNA, RNA, protein) contribute to disease progression and treatment response. Ex. it is often associated with poor prognosis.
- Singel-cell targeted DNA-seq: WGA, multiplex PCR to amplify 90 SNVs and CNAs (using SNPs). Validate the result using cancer cell fraction (CCF), comparing with WES estimates. Some patients show clearly more than one clones. Ex. one patient, one subclone has deletion in 17p that contains p53, and another TP53 mutation.
- ScRNA-seq study: first identify gene sets that show overdispersion across cells. Then for each cell, show the PC scores of each gene set (representing the activity of a gene set in each cell). Found a few processes not found in mutation studies: e.g. antigene presentation.
- Single-cell targeted RNA: cDNA pre-amplification, then DNA and qCPR for mRNA of 96 genes. Analyze 384 cells. Genes with low expression in bulk RNA: show bimodal expression.
- Joint DNA-RNA analysis: reconstruct phylogeny from DNA data, and consistent with scDNA analysis. In one patient, genes identified from scRNA-seq analysis that drive transriptional heterogeneity do not show significant expression difference in two subclones.
- Identify LCP1 as a new CLL driver: the mutations that are common in a major subclone is likely a driver mutation.
- Hypothesis: convergent evolution in tumor, at DNA level: TP53 mutation may evolve independently in two subclones. At RNA level: different subclone evolve similar transcriptome/phenotypes, even though genetically distinct.
- Lessons: (1) Bimodal expression at single-cell level, esp. for low-expression genes. (2) Possible to identify driver mutations from cell phylogeny. (3) No simple relationship between genotype and transcriptome (phenotype).

Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer [Chung and Park, NC, 2017]

- Background: why heterogeneity matters? Ex. in ER+ breast cancer, some of the cells may be ER-, and this may explain drug resistance.
- Data: 515 scRNA-seq from 11 breast cancer patients, covering 4 major subtypes. Also WES data.
- Somatic mutations: about 500 mutations per sample at AF ¿ 0.03, but only 100 for coding mutations.
- Validation of scRNA-seq data: RT-PCR of selected genes. PCA and how samples are clustered by patients or tumor subtypes.
- Define carcinoma and non-cancer cells in scRNA-seq: use chromosomal expression, show that it matches CNVs. Use these patterns to classify tumor and non-tumor cells. For non-tumor cells, study their immune or stromal signature (know gene sets). Results: 317 epithelial breast cancer cells, 175 immune cells and 23 non-carcinoma stromal cells.

Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia [Nature, 2019]

- CLL patients, sc-RRBS. Use epi-mutations (simply compare reads) to reconstruct the linear tree: use binary methylation status, ML tree.
- CLL tree: early branching, higher epi-mutation rates, comparing with controls.

Chapter 6

Cancer Immunology

Background for cancer immunology [personal notes]

- How T cells regulate and kill tumor cells? (1) IFN-γ, from Th1 cells, stop cell proliferation. (2) Induce apoptosis by FasL/Fas and TRAIL/TRAIL-R interactions.
- Cytokines that control T cells: (1) Activation: IL-12, for Th1 cell activation; IL-6, for T cell differentiation. (2) Repression: IL-10, TGF-β.
- Types of Macrophages: (1) M1 MΦ: immune protection against pathogens. Activated by IFN-γ from Th1 cells, and generate pro-inflammatory cytokines such sa IFN-γ, IL-12, IL-16, TNF, IL-1. (2) M2 MΦ: wound healing and tissue repair. Activated by IL-10, IL-13, IL-4. (3) Tumor associated macrophages (TAM): promote tumor cell proliferation, metastasis and angiogenesis. Secrete IL-10 and TGF-β.

Mechanisms of immune elimination and escape [personal notes; Mittal and Smyth, COI, 2014]

- Immune elimination (FIgure 1 of Mittal)
  - Expression of stress ligand and DAMP molecules, activation of innate immunity
  - MHC expression and antigen presentation, activation of CD8 T cells.
  - Expression of NKG2D ligand, activation of NK cells.
  - IFN-γ by CD4, DCD8 T cells and macrophages: stop tumor proliferation.
- Immune escape (Figure 2 of Mittal):
  - Escape from recognition: reduced expression of MHC.
  - Escape from killing by CD8 T and NK cells: upregulation of STAT3, Bcl2, and down-regulation of Fas and TRAIL-R.
  - Angiogenesis: expression of VEGF.
  - Suppress CD8 T cells: IDO, TDO, PD-L1, CD39, CD73.
  - Help by suppressive immune cells, including MDSCs, M2 macrophages and some DCs: secretion by these cells IL-10 and TGF-β, which CD8 T cells and NK cells.

Pathways of cancer-immune cell interactions [personal notes]

- Tumor associated inflammation: IL-1 beta, IL 18 promote inflammation.
- Antigen presentation: UPR response can down-regulate MHC I expression, and reduce antigen presentation.

- Checkpoint:
- T cell priming/activation: CD40.

Questions about anti-tumor immune response [personal notes]

- What's the role of senescence in cancer? It is supposed to be a cancer protective mechanism (no cell division), but also triggers inflammation.
- Immune and inflammatory responses are generally resolved after pathogens are eliminated. What are the mechanisms? Do tumors hijack this system?
- How do organsims without adaptive immune system control cancer? Difficult problem because cancer cells are similar to normal cells.

Cancer immunoediting: integrating immunitys roles in cancer suppression and promotion [Schreiber and Smith, Science, 2011]

- Experiment proving immunosurvellience: knockout mice with no IFN-gamma response or no adaptive immunity (RAG2) are much more susceptible to tumor.
- Experiment proving immunoediting: immuno-deficient mice, take tumors and inject to WT mice. About 50% are rejected because of high immunogenicity, showing that immune system can edit the tumor antigens.
- Possible sources of tumor neoantigens: somatic mutations, aberrantly expressed genes, overexpressed genes (e.g. HER2), CT antigens.
- Elimination phase: triggers for innate immunity include: classical damage signal such as type I IFN, stress ligands, damage-associated molecular pattern molecules (DAMPs) from dying tumor cells or damaged tissues.
- Equilibrium phase: immune system keeps tumor cells in functional dormancy via both cytocidal actions and inhibitory signal.
- Escape phase: reduced antigen presentation (and elimination of immunogenic clones); immunosuppressive cytokines such as TGF-beta, IL-10, IDO, galectin; inhibitory immune cells, most importantly T-reg and MDSCs (acting via cytokines).
- Inflammation: chronic inflammation promotes tumor by multiple mechanisms: genotoxic stress, cellular proliferation (wound healing?), angiogenesis and tissue invasion. Paradoxical roles of Myd88 and IL-1beta, TNF-alpha, IFN-gamma, both tumor-promoting and anti-tumor.
- Future direction: how types of neoantigens are determined by cellular transformations? Whats the mechanism of equilibrium?

New insights into cancer immunoediting and its three component phases elimination, equilibrium and escape [Mittal and Smyth, COI, 2014]

- Background: immunodominance, immune responses are mounted against only a few of the antigenic peptides out of the many produced.
- Immunoediting: the role of immune system in shaping the immunogeneicity of tumor (immune system edits tumor). Generally, this process depends on the degree of immunocompetence of the host.
- Somatic mutations as tumor antigens and role of adaptive immune system: (1) Evidence: in a mouse model, identifies a point mutation that acts as a major immunodominant rejection antigen. Outgrowth of clones lacking this antigen. (2) Exome analysis can identify tumor-specific mutational antigens. (3) Negative selection against tumor antigens: in cancer vaccine studies, need of using multiple tumor antigens.

- Evidence of innate immune system: NK cells, activated by local IL-12, can activate M1 macrophage in tumor.
- Elimination process (Figure 2): (1) Triggering innate immunity, esp. DCs: tumor cell expression of stress-related genes, NKG2D ligands, type 1 IFN- response. (2) Activation of adaptive immunity by DCs: CD8 T cells, CD4 T cells and NKT. (3) Response of adaptive immune system: killing by CD8 T cells, IFN-gamma which inhibits proliferation. (4) Role of other players of innate immune system: Macrophages that secrete IFN-gamma, and granulocytes which secrete TNF-alpha, IL-1.
- Cell ploidy changes can be monitored by immune system, via T cells and INF-gamma.
- DNA damage response: e.g. induced by Myc expression, can enhance immune recognition.
- Senescent tumor cells can also induce immune elimination: expression of p53, secrete cytokines (IL-6, IL-12) that recruit NK cells.
- Equilibrium: balance between Elimination and Escape. Cell level: comparing with Escape, more CD8 T, NK, gamma-delta T, and fewer T-reg, NKT and MDSC. Cytokine level: IL12 and IFN-gamma vs. IL-10, IL-23.
- Escape process (Figure 4).
- Immune signatures that are predictive of outcome: (1) Positive: infiltrating Th1, CD8 T cells. (2) Negative: myeloid cells, Th17.
- Impact of microbiota: microbial products may trigger inflammation, which drives tumor growth.
- Remark: is immunocompetence of the host affected by germline variants? If so, we would expect interactions between germline and somatic mutations.
- Q: NKG2D is induced in tumors. Whats the trigger?
- Q: How are DCs activated in tumor? Possible: type I IFN response in host cells.

Immunotherapy of cancer [Elaine Mardis, Jun, 2016]

- Designing cancer vaccine:
  - Prediction of neoantigens: prediction of mutational peptides and MHC binding.
  - Neoantigen selection: use both prediction program and expression of neoantigen from RNA-seq data.
  - Dendritic cell (DC) vaccine: Load DCs with neoantigen peptides, then use them to stimulate CD8+ T cells.
- Questions/directions in immunotherapy:
  - Better prediction of neoantigens: frameshift indels, rare HLA haplotypes, MHC class II prediction.
  - Vaccine platform: how to delieve vaccine, DNA, or RNA or protein. Note: only peptide can load MHC, but we can may mRNA to more efficiently delierve peptides.
  - Combination of vaccine and check-point inhibitor.
- Prediction of response to immunotherapy: should combine the immuno-suppressive potential of cancer and neoantigen load. Current practice: PD-1 level via antibody staining.
- Questions:
  - How does body tolerate normal somatic mutations?

– How to use the knowledge of level of neoantigens? The ones very early in the cancer evolution should be present more often, and more effective.

Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity [Rooney and Hacohen, Cell, 2015]

- Measuring cytolytic activity (CYT): using granzyme and perforin levels. CYT activity varies across tumor types: in some tumors, e.g. kidney and cervical cancer, strong induction; but in others, e.g. breast cancer, no.
- Correlated genes with CYT: other genes that are markers of CYT activity; also positive correlation with genes usually not expressed in CTLs and NK cells, some are immunosuppresive molecules. Explanation: CYT induces tumor expression of immunosuppressive genes.
- Neoantigen load: CYTs are positively correlated with mutation load and neoantigen load. Also find depletion of neoantigen in colorecetal and kidney cancer, comparing with synonymous mutations.
- CYT also correlates with virus, ERVs and ecotopic gene expression (cancer testis genes).
- Mutations associated with CYT: regression controlling for tumor type and background mutation rates. 35 genes, except TP53, all positive correlation, i.e. higher mutations in CYT active tumors. Genes are involved in (1) Antigen presentation: B2M and MHC I. (2) Escape: CASP8. (3) CT antigens. (4) Innate immune sensing: DDX3X, ARID2.
- CNVs associated with CYT: PDL1/L2, amplification positively associated with CYT. IDO1/2 and ALOX12B/15B (potent immunosuppressor): amplification in low CYT tumors.
- Model: (1) Neoantigens, virus/ERVs: triggers CYT. (2) Emergence of evading subclones: B2M, HLA, CASP8 LoF mutations, and PDL1/L2 upregualtion. (3) Emergence of suppressive subclones (non-autonamous mechanism): TP53, ALOX, IDO1/2.
- Remark: induction of CYT may also be influenced by innate immune sensing.
- Remark: the analysis of mutation-CYT association is largely about positive selection in high CYT tumors.

The Immune Landscape of Cancer [Thorsson, Cell, 2018]

- Methods for defining immune subtypes: (1) collect 160 expression signatures: gene sets, PCs from gene signatures. (2) Choosing representative signatures: WGCNA on signatures, then find modules from WGCNA, then find the signatures that correlate with eigen-signatures for the modules. In the end, 5 signatures, and clustering tumor sample expression data.
- Biological characteristics of 6 subtypes: the six clusters have a lot of overlap on the 5 signatures. Notable features of the six clusters: (1) Wound healing: angiogenesis and high proliferation. (2) IFN-gamma: low macrophages, highest CD8 T cells, and high proliferation. (3) Inflammatory. (4) Lymphocyte depleted: minimum lymphocyte, but normal INF-gamma, high macrophages. (5) Immunologically quiet: low lymphocyte, low INF-gamma. (6) TGF-beta: high TLI.
- Composition of tumor cell infiltrate: use CIBERSOFT to estimate cell proportions (22 reference transcriptomes). Leukocyte fraction (LF) varies across tumor types: 1 - 20% median.
- Prognostic associations: train an elastic net regression model using cell compositions (LF, CIBERSOFT), transcriptome signatures (6 subtype scores, etc.), neoantigen (mutations, testis antigen expression) and TCR/BCR diversity. Associated with better prognosis: lymphocyte, TCR diversity and cytokine from Th1 and Th17 cells; with poor prognosis: macrophage, wound healing, TGF-beta.

- Association of somatic mutations with 6 subtypes: contingency table of mutations, subtypes. Fishers exact test. Found 31 significant associations (Figure 4c), mostly driven by increased mutations in C1 (wound healing) and C2 (IFN-gamma). C5: 3 enriched genes IDH1, ATRX (chromatin) and CIC (transcription repressor). Note: The test is not comparing mutations with expected, in contrast, if for some subyptes, mutations are enriched, they must be depleted in some other subtypes.
- Association of somatic mutations with LF (Figure 4D): linear regression of LF with mutations, controlling for tumor type and total mutation count in a patient (pan-cancer analysis). Only test on known driver genes, and likely functional mutations. 30 genes have FDR ¡ 0.1. 15 genes are enriched in low LF patients: IDH1, CIC, FGFR3, GATA3, NRAS, KRAS. Most are found in pan-cancer, some tumor type specific. For most signaling pathways, PI3K, NOTCH and Ras, mutations associate with LF in different directions in different tumors, only TGF-beta show consistent effects.
- Lesson: clustering tumor immunoprofiles, first do dimension reduction using gene signatures (e.g. modules from WGCNA, or predefined gene sets), then cluster.
- Remark: the 6 subtypes heavily overlap in terms of biological characterisitics.
- Remark: the regression analysis of LF and genes: possibly confounding. Often for the classical driver genes/pathways, they are associated to LF with different directions in different tumors.

Chromatin regulation and immune escape [Science, 2018]

- Background: type I IFN response: IFN-alpha and IFN-beta, triggered by viral infection. Type II INF response: IFN-gamma (from T-cells, esp. Th1 cells), lead to MHC expression, and apoptosis - generally expression of tumor suppressor genes. IFN-gamma downstream: activation of STAT1, which activates ISG (interferon-stimulated genes) expression.
- Negative regulator of IFN-gamma: LoF mutation of PBRM1 makes tumor cells sensitive to checkpoint blockade (CRISPR screening assay).
- Other cancer genes may act via regulating ISG genes: SWI/SNF, EZH2 (subunit of PRC2).
- **Lesson**: The general lesson is: in vertebrates, the decision of a cell (e.g. apoptosis) is controlled not only by internal conditions such as DNA damage and developmental fate, but also by the immune system. IFN-gamma pathway is a key link between the immune system and tissues cells (signal/command other cells to express MHC or commit suicide). Dis-regulation is involved in cancer.

Dysregulated IL-18 Is a Key Driver of Immunosuppression and a Possible Therapeutic Target in the Multiple Myeloma Microenvironment [Cancer Cell, 2018]

- Background: DAMP (damage-associated molecular pattern) is recognized by pattern recognition receptors (PRRs), e.g. TLRs. This leads to tumor-promoting inflammation, which mobilizes MDSCs and TAMs.
- Background: IL-18 is IL-1 family cytokine.
- Role of IL18: mice deficient in IL18 shows protection from Multiple myeloma. Mechanism: IL18 activates MDSCs, which promotes cancer progression.

The Immune Revolution: A Case for Priming, Not Checkpoint [Cancer Cell, 2018]

- CD40 is a receptor (TNF family) expressed in DCs, B cells. CD40 is important for activation of DCs: upregulation of cytokines (such as IL12), antigen presentation, etc.
- CD40 agonist: new immunotherapy strategy, activate DCs and T-cell priming.

How dormant cancer persists and reawakens [Science, 2018]

- Disseminated cancer cells (DCCs): can evade immune system. How?
- Model: DCCs activate UPR (in response to internal or external stress), which leads to down-regulation of MHC I, escaping detection of CD8+ T cells.
- Q: Why stress response leads to MHC I down-regulation? Hypothesis: stress condition may interfere with antigen presentation. More likely to present something perceived as foreign by CD8+ T cells (e.g. protein misfolding), so down-regulate antigen presentation to avoid cell death.

Tumor-derived TGF- inhibits mitochondrial respiration to suppress IFN- production by human CD4+ T cells [Science Signaling, 2019]

- TGF-beta can promote tumor by increasing angiogenesis.
- TGF-beta may suppress anti-tumor activity by: impair mitochrondrial activity of CD4 T cells, and block production of IFN-gamma.

Blockade of EGFR improves responsiveness to PD-1 blockade in EGFR-mutated nonsmall cell lung cancer [Science Immunology, 2020]

- In EGFR mutated NSCLC: usually low TMB, and low CD8 T cell infiltration and high T-reg cells.
- EGFR mutated line vs. cell lines with no EGFR mutation: reduced expression of chemokines/cytokines that recruit CD8 T cells.
- EGFR affects immune genes by changing cJun and IRF.
- **Remark**: What is the rationale of the connection between EGFR and immune genes (cytokines/chemokines)? Hypothesis: EGFR is used in wound healing, which suppress CD8 T cell infiltration, potentially through acting on T-reg cells. Ref: Cutting Edge: Regulatory T Cells Facilitate Cutaneous Wound Healing [J Immunology, 2016]

New predictors for immunotherapy responses sharpen our view of the tumour microenvironment [Nature, 2020]

- B cell function in anti-tumor immunity: sometimes inhibitory of T cells; sometimes neo-antigens activate BCRs so that B cells generate antibody. Antibodies can tag cancer cells, leading to attack by other immune cells (antibody-dependent cell death) and can also educate T cells.
- Tertiary lymphoid structures (TLS): clusters of B cell and T cells (outside). Associated with favorable outcome of immunotherapy.

6.1 Adaptive Immunity in Cancer

Landscape of tumor-infiltrating T cell repertoire of human cancers [Bo Li and Xiaole Liu, NG, 2016]

- Background: TCR has alpha and beta chains, similar to light and heavy chains, respectively. No somatic hypermutation. VJ recombination for light/alpha; and VDJ for heavy/beta. Junctional diversity (between V and J): CDR3, about 20-30 AAs between V and J, chosen from a much longer sequence (multiple smaller segments ¡10 nt), involved in antigen recognition.
- Background: PD-L1 expression: normally expressed in APCs (should not be targets of cytotoxic T cells) and helper T cells (used to turn off other T cells).
- CDR3 assembly: use PE reads, if one part is mapped to V, the other unmapped likely from CDR3. Grouping unmapped reads, and do de novo assembly. Each tumor sample has multiple assembled TCRs. One sample: 224M reads ¿ 52K reads in TCR regions ¿ call 56 CDR3 sequences.

- Building TCR profiles of tumor samples: use of V segments (among 50): 50-dimension vector for TCRs (average profile per sample). PC plots: show difference with tumor types.
- Gamma-delta T cell faction: different across tumor types. Two types of TCRs: alpha/beta, gamma/delta - may also be involved in innate response.
- CDR3: length variable from 6 - 25 AAs. Motif logo: highly variable in the middle, but conserved across TCGA samples (Figure 2). Comparison with healthy individuals: share very little. Q: ascertainment bias? Explanation?
- Relationship between TCR diversity and mutational load: use CPK to measure diversity (one CDR3 region: one TCR clone).
- CT antigens (expression trigger immune response - normally expressed in cancer): expression of two genes positively associated with CPK. Significance: could be used as vaccines.
- Co-occurence of somatic mutations and CDR3 motifs: find 3 significant pairs. One mutation: good binding with MHC.
- Q: Why CDR3 in cancer shows little variation? Is this expected?
- Remark: difficult to interpret TCRs, e.g. do not know what antigens (what genes) may trigger which TCRs.
- **Lesson**: representation of TCRs, e.g. what V or J fragment is used? CDR3 sequence motifs. TCR repertoire may reflect tumor types, mutation load, specific somatic mutations, CT antigens and T cell subtypes.

Landscape of B cell immunity and related immune evasion in human cancers [Hu and Liu, NG, 2019]

- Background: B cells when seeing antigens: somatic hypermutations and class-switch recombination. Ig class switching: IgM ¿ IgG involves removal of certain regions. Class switch usually indicates B cell function change, e.g. maturation.
- Use TRUST to extract the B cell immunoglobulin hypervariable regions from bulk tumor RNA-sequencing data. CDR3 mapping: for unmapped reads, if one pair can be mapped to VDJ, clusters of reads with one end matching. Obtain disjoint cliques, then contigs.
- Grouping B cell clones: allow mutations (somatic hypermutations). Construct the tree of BCRs.
- Ig class switching: from BCR clusters, infer the Ig classes. Find that in the same clusters: multiple Ig classes. Found IgG1 to 3 switch: among all BCR clusters, half of them have both IgG1 and 3.
- Comparison of BCR in tumor and adjacent normal: infiltration, diversity, IgA switch, only IgG switch differs significantly.
- Prevalent somatic copy number alterations in the MICA and MICB genes (MHC I related) related to antibody-dependent cell-mediated cytotoxicity were identified in tumors with elevated B cell activity.
- **Model** (Figure 6c): ADCC (antibody mediated cancer cell cytotoxicity) via NK cells. Evidence that IgG switch correlates with NK cell immune escape. Q: IgG class switch is promoted by cancer?

Chapter 7

Cancer Diagnosis and Treatment

RNA-seq for blood-based pan-cancer diagnostics [NRG, 2016]

- Principle: blood platelets change their mRNA expression profile in response to contact with a tumour (so-called 'tumour-educated platelets').
- Platelet samples in controls vs. metastatic patients: Selection of a classifier-specific gene list of 1,072 RNAs. Prediction accuracy close to 95%. Also 89% accuracy of predicting tumor types.
- Limitation: inability to discriminate between different cancer stages and between metastasized versus non-metastasized tumours.

New single cell methods for oncology and immuno-oncology [Jim Heath, ISB, 2020]

- I. Cancer drug resistance.
- Melanocyte ¿ mesenchymal: de-differentiation process. BRAF inhibitor or immunotherapy can block the process. Chromatin state changes triggered by BRAFi.
- Single-cell level protein: detection antibodies in slides. Measure protein changes in first few days of treatment. Analysis: t-SNE analysis (or psuedotime), show that cells from the same days are clustered.
- Cell state changes: glucose take initially high, but reduced in next few days. MITF marker: changes.
- Trajectory analysis from Day 0 to Day 5: likely 2 distinct paths (depletion of cells in between).
- Two paths: (critical points) distinct genes, different druggable genes. Which trajectories specific cells take: depend on pre-existing states, e.g. TFs. Note: these are isogenic cells.
- Lesson: drug treatment may change differentiation path of tumor cells (independent of genetic changes).
- II. CD8 T cells: target ALG8. Goal: pair TCRs with neoantigens.
- Adoptive Cell therapy: isolation of tumor reactive T cells; then do cloning and expansion; and back to patients.
- Procedure: define neoantigens from tumor (bioinformatic predictions). Using nanoparticles to present pMHC (peptide-MHC) and DNA oligonucleotides. Then treating T cells and T cell clonal expansion.
- Single chain pMHC molecules: single molecule of MHC, neo-epitope, B2M. Use UV exchange. Q: what is it?
- Issue: loading of neoantigen peptide to MHC is rate limitation (free energy barrier).
- Future: CRISPR knock-in of anti-tumor TCRs. Enginner resistance against immune checkpoints.