REVIEW

doi:10.1038/nature10762

Clonal evolution in cancer

Mel Greaves[1] & Carlo C. Maley[2]

Cancers evolve by a reiterative process of clonal expansion, genetic diversification and clonal selection within the adaptive landscapes of tissue ecosystems. The dynamics are complex, with highly variable patterns of genetic diversity and resulting clonal architecture. Therapeutic intervention may destroy cancer clones and erode their habitats, but it can also inadvertently provide a potent selective pressure for the expansion of resistant variants. The inherently Darwinian character of cancer is the primary reason for this therapeutic failure, but it may also hold the key to more effective control.

Cancer is a major cause of death throughout the world and, despite an extraordinary amount of effort and money spent, the eradication or control of advanced disease has not been achieved[1]. Although we have a much greater understanding of cancer biology and genetics[2], translation into clinical practice needs to allow for the cellular complexity of the disease and its dynamic, evolutionary characteristics. These features provide both barriers to, and opportunities for, successful treatment.

In 1976, Peter Nowell[3] published a landmark perspective on cancer as an evolutionary process that is driven by stepwise, somatic-cell mutations with sequential, subclonal selection. This is a parallel to Darwinian natural selection, with cancer clones as the equivalent of asexually reproducing, unicellular quasi-species. Modern cancer biology and genomics have validated cancer as a complex, Darwinian, adaptive system[4,5] (Box 1 and Supplementary Information).

Cancer-clone evolution takes place within tissue ecosystem habitats. These habitats have evolved over a billion years to optimize multicellular function but restrain clonal expansion of renegade cells. However, the resilience of multicellular and long-lived animals depends on the phenotypic properties that, if not tightly regulated, drive or sustain malignancy: that is, cellular self-renewal and stabilization of telomeres, which allow extensive proliferation, angiogenesis, cell migration and invasion[6].

The long time period usually required for cancer symptoms to emerge and the complexity of the resultant mutations is, in part, a reflection of the sequential and random searches for phenotypic solutions to constraints from the micro-environment. The evolutionary progression of cancer is usually stalled or aborted, as shown by the high frequency of clinically covert premalignant lesions[7–9]. Cancer-suppressive mechanisms relegate most cancers to old age, when they have little effect on the reproductive fitness of their hosts.

Limited resources, environment architecture and other constraints of the micro-environment limit the size of solid tumours at every stage of their progression. Even advanced malignancies can show Gompertzian growth[10] — the cancer cell doubling time (around 1–2 days) is orders of magnitude faster than tumour doubling time (around 60–200 days)[10] — implying that the vast majority of cancer cells either die before they can divide[11] or are kept from dividing by the tumour micro-environment. Thus, natural selection in tumours, in the same way as selection in organisms, takes place through competition for space and resources.

Oncologists change cancer-clone dynamics by introducing a potent source of artificial selection in the form of drugs or radiation, but evolutionary principles still apply. Usually, treatment will result in massive cell death, which provides a selective pressure for the proliferation of variant cells that resist treatment (the mechanisms for this are discussed later). Furthermore, many cancer therapeutics are genotoxic; cells surviving treatment, which could then go on to regenerate the cancer, may have mutated further, resulting in cells with improved fitness and malignant potential.

The tools of and insights from evolutionary biology and ecology can therefore be applied to the dynamics of cancer before and after treatment to explain the modest returns from cancer therapy. We show that cancer is an inherently evolutionary process and suggest alternative strategies for effective control.

Mutational drivers and clonal dynamics

The basic principle of a Darwinian evolutionary system is the purposeless genetic variation of reproductive individuals who are united by common descent, together with natural selection of the fittest variants. Cancer is a clear example of such a system. Most mutational processes have a bias at the DNA sequence level. The particular mutational spectra in a cancer cell can be a reflection of error-prone repair processes or associated with a genotoxic exposure (for example, cigarette carcinogens, ultraviolet light and chemotherapeutic drugs[2]). The patterns of genetic instability (chromosomal or microsatellite) in cancer cells may reflect exposure to, and the selective pressure exerted by, some classes of chemical carcinogens[2]. Nevertheless, for the functions encoded in genes, mutagenic processes are essentially blind or non-purposeful (with the exception of intrinsic mutagenic or recombinatorial enzymes preferentially targeting lymphoid immunoglobin or T-cell receptor genes[12]). The recurrent, mutation-endowed fitness traits in cancer reflect the potent impact clonal selection can have.

Clones evolve through the interaction of selectively advantageous 'driver' lesions, selectively neutral 'passenger' lesions and deleterious lesions (a 'hitchhiker' mutation in evolutionary biology is equivalent to a passenger mutation in cancer biology). In addition, 'mutator' lesions increase the rate of other genetic changes[13,14], and micro-environmental[15] changes alter the fitness effects of those lesions. The identification of driver lesions is supported by the independent observation that these lesions occur more frequently in multiple neoplasms than would be expected in the normal background mutation rate, that they are associated with clonal expansions[16,17] and from the type of mutation seen (missense, nonsense, frameshift, splice site, phosphorylation sites and double deletions)[18–20], particularly if the gene involved has a known role in cellular processes relevant to oncogenesis. The evidence gained from genetic studies

[1]Division of Molecular Pathology, The Institute of Cancer Research, Brookes Lawley Building, 15 Cotswold Road, Sutton, Surrey SM2 5NG, UK. [2]Center for Evolution and Cancer, Helen Diller Family Comprehensive Cancer Center, Department of Surgery, University of California, 2340 Sutter Street PO Box 1351, San Francisco, California 94115, USA.

of human tumours should be corroborated with functional tests and animal models. Passenger lesion status can also be ambiguous or context-dependent: for example, cases of monoallelic loss that only impact on function when the second allele of the same gene is lost, mutations that only cause a phenotypic effect when another gene locus also mutates, or cases in which the mutants are functionally relevant only in the context of therapeutic responses involving that gene.

Only a few studies have attempted to quantify the selective advantage provided by driver mutations. Bozic *et al.*[21] (using a non-spatial population genetics model of sequential, exponential clonal expansion) derived a formula for the proportion of expected neutral passenger mutations versus the proportion of selectively advantageous driver mutations as a function of the selective advantage of the driver mutations. By fitting this equation to glioblastoma and pancreatic cancer resequencing data, the authors estimated that driver mutations gave an average fitness advantage of only 0.4% (ref. 21). To measure the mutant clone selective advantage directly would require longitudinal samples of a neoplasm and estimation of the clone sizes at each time point.

The dynamics of somatic evolution depend on the interaction of mutation rate and clonal expansion. Mutation rate varies substantially between different genomic regions[22] and between different types of abnormality (for example, single-base sequence changes versus balanced chromosomal rearrangements and gene fusions), and mutation rates will increase by the instigation of genetic instability[23–25]. The rate of epigenetic change has been estimated to be orders of magnitude higher than that of genetic change[26], and could be a major determinant of clonal evolution. Natural selection affects epigenetic variation within neoplasms[27], because epigenetic changes are inherited at cell division and can affect cell phenotypes. Evolutionary biology tools to address many of these mutation rate complexities exist (see Supplementary Information), but these remain under used in cancer biology[28]. The traditional model of clonal evolution suggests that a series of clonal expansions grows to dominate the neoplasm ('selective sweeps')[16,21,29], but this can occur only if the time to the next driver mutation is longer than the time required for a clone to sweep through the neoplasm. In addition, if the second mutation occurs in a competitor clone, the expansion of both clones is restrained by mutual competition (known as clonal interference)[30]. Given the large population size and high mutation rate typical of neoplasms, clonal competition is probably common[31,32]. This issue is best addressed by serial sampling, and limited data suggest that parallel clonal expansions occur before subclones begin to dominate in early cancer development[33–35]. Initial evidence indicates that large clonal expansions after cell transformation are rare[26]. Direct evidence, from serial sampling of oncogenic mutations in advanced disease[36], metastasis[37] or post-chemotherapy relapses (see Supplementary Information), indicates selective sweeps originate from pre-existing genetic variants or subclones.

Punctuated equilibrium versus gradualism

The argument of gradualism versus punctuated equilibrium[38] (a longstanding debate in species evolution) has recently emerged in the consideration of the clonal evolution of neoplasms. It is unknown whether malignant clones, with their markedly altered genomes, evolve gradually through a sequence of genetic alterations and clonal expansions; accumulate many lesions over time in a rare, undetected subclone that finally appears in a clonal expansion; or have a few, large-scale punctuated changes, possibly prompted by an acute insult or a single, catastrophic mitotic event that generates multiple lesions across the genome (or on a single chromosome, known as chromothripsis)[39]. Evidence of tens of non-synonymous mutations in cancers was interpreted under the assumption that they were generated by tens of clonal expansions[29]. Reconstruction of genealogies of neoplastic clones, based on genetic heterogeneity within neoplasms, suggests that clones with ancestral genomes are not driven to extinction by later clonal expansions[31–33], which allows the history of a neoplasm to be revealed. Breast cancer data[32] have shown that clones with intermediate genotypes are difficult to detect; each clone generates a cloud of genetic neutral or non-viable subclone variants around it. A study of B-cell chronic lymphocytic leukaemia[40] suggests that intermediate clones can be detected, but at a frequency of <0.001, which was below the detection threshold of the breast cancer study[32]. Intermediate clones may be rare because they have had limited potential to expand or because they were once common but were outcompeted by more recent clones.

The frequency of premalignant clonal lesions (or carcinoma *in situ*) substantially exceeds clinical cancer rates[7–9]. This, as well as cancer dormancy[41] and genetic reconstitution of clonal histories[37], indicates that cancer clones have long periods of stasis. However, cancer-clone evolution probably passes a point of no return, possibly at the metastatic growth stage. If unlimited proliferative capacity is guaranteed by telomere stabilization[25], then clonal expansion is stopped only when the size threatens the life of the patient. When provided (albeit rarely) with the routes for dissemination and immunoselection, cancer cells can have a parasite-like immortality and can re-establish themselves in other individuals[6,42,43].

BOX 1

Cancer as a complex system

- Cancers exist in a variety of taxonomic quasi-classes, genera, species, characterized by divergent cells of origin and mutational spectra. Each cancer is unique.
- Cancers evolve over a variable time frame (anywhere from 1 to 50 years), and the clonal structure, genotype and phenotype can shift over time in each patient. Each cancer is, in effect, multiple different (subclonal) cancers that occupy overlapping or distinct tissue habitats.
- The number of mutations in a cancer can vary from a handful (10–20) to (the more usual) hundreds or thousands. The great majority are passengers, and a modest, but undefined, number are functionally relevant drivers. The mutational processes are very diverse.
- Cancers acquire, through mutational and epigenetic changes, a variety of phenotypic traits that compound to allow territorial expansion, by proliferative self-renewal, migration and invasion — properties that are cruical to normal developmental, physiological and repair processes.
- Advanced, disseminated or very malignant cancers seem to be almost uniquely competent to evade therapy.
- Most, if not all, of this complexity can be explained by classical evolutionary principles.

The cancer ecosystem

Tissue ecosystems provide the venue and determinants for fitness selection (the adaptive landscape[44]). Tissue micro-environments are complex, dynamic states with multiple components that can influence cancer-clone evolution (Fig. 1). For example, transforming growth factor-β is a cancer-ecosystem regulatory molecule[45]. Other cellular and cytokine components of inflammatory lesions are potent and common modulators of the cancer-cell ecosystem[25].

The interaction between cancer cells and their tissue habitats is reciprocal. Cancer cells can remodel tissue micro-environments and specialized niches to their competitive advantage[46]. Cancer-clone expansion is controlled by architectural constraints or barriers, such as sequestration of stem cells into crypts in the gastrointestinal tract[47]

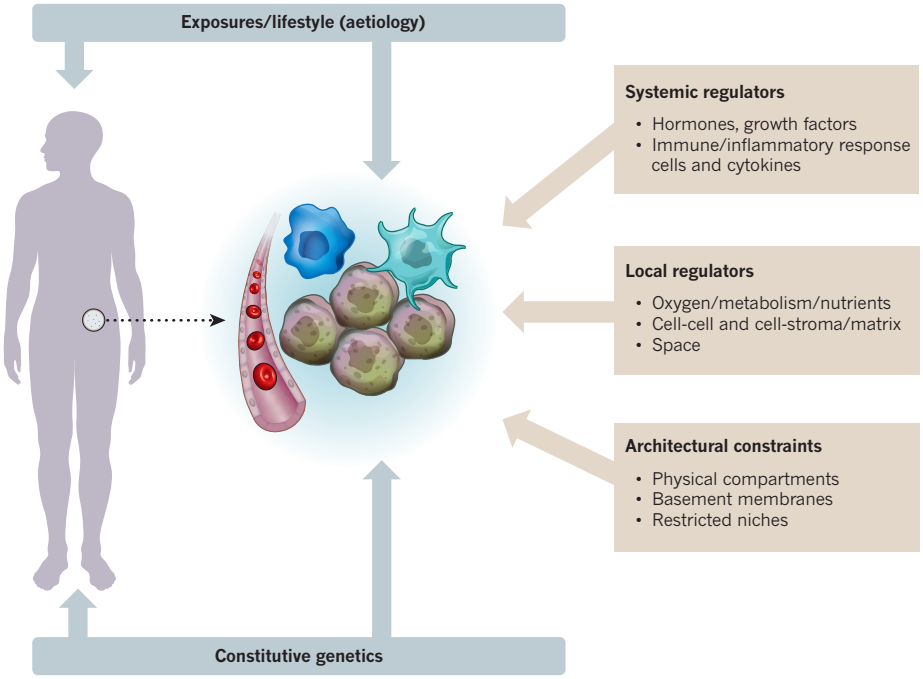

Figure 1 | The complexity of tissue ecosystems. Exposure, the constitutive genetics of the host cells, systemic regulators, local regulators and architectural constraints all impinge on the evolution of somatic cells.

and the need for external signals for proliferation and cell survival. However, some micro-environmental components can promote neoplastic cells; for example, infiltrating macrophages and neovascularization, in response to anoxia, can support neoplastic cell survival and proliferation. Mathematical modelling shows cancer-clone evolutionary selection for more robust or malignant phenotypes is less likely in more stable or homogeneous micro-environments[48]. Spatial heterogeneity of resources in the primary tumour selects for cell migration and emigration, which may explain why there is selection for metastasis[49]. Preclinical models have suggested that normalizing the resources across the primary tumour can suppress metastasis[50]. As clones and subclones expand, migrant cells invade new habitats within and between tissues, in which they experience new selective pressures that can cause further cancer-cell diversity. This malignant feature, and its associated morbidity, characterizes end-stage cancer.

Cancer-cell habitats are not closed systems. The tissue ecosystem, in addition to regulation by systemic factors (such as nutrients and hormones) or invasion by inflammatory or endothelial cells, is modified by external factors. As well as the tissue site, the ecosystem for each cancer includes environmental, lifestyle and associated aetiological exposure of the patient. Genotoxic exposure (such as, cigarette carcinogens or ultraviolet light), infection, and long-term dietary and exercise habits that affect calorie, hormone or inflammation levels can have a profound effect on the tissue micro-environments, as well as directly on cancer cells (Fig. 1). These factors are the aetiological link to the initiation or progression of cancer, and without such modulating exposure, the risk of cancer-clone initiation and evolution would be reduced.

Cancer-tissue ecosystems can be radically altered after chemotherapy or radiotherapy. Most cancer cells may be decimated, but the remodelled landscape creates new selective pressures, resources and opportunities that may allow pre-existing variant cancer cells that survived treatment to emerge. Crucially, stroma or specialized habitat niches may protect cancer cells against the therapy[51].

Cancer genomics and clonal architecture

Cancer-genome sequencing, facilitated by the introduction of second-generation whole-genome sequencing, has provided further insight into the complexity of the genetics and evolutionary biology of cancer cells[2]. In most cases, transformation and metastases are probably clonal[2] because they are derived from single cells; therefore, the identification of the mutations present in all of the cells of a tumour can help to reconstruct the genotype of the founder cell. These founder events limit the genetic and clonal complexity of tumours. We already had a long list of recurring driver mutations (with gain or loss of function) as a result of the fine mapping of chromosomal breaks, candidate gene sequencing and functional screening of bulk samples from tumours. However, the use of genomic screens has demonstrated the scale of cancer-genome complexity. Individual cancers can contain hundreds, or tens of thousands, of mutations and chromosomal alterations[2]. The great majority of these are assumed to be neutral mutations arising from genetic instability. Chromosomal instability (amplifications, deletions, translocations and other structural changes) is a common feature, but it is not clear whether there is an increased rate of simple base-pair mutations in cancer[2,21,23,52]. Evolutionarily neutral alterations are thought to register in the screens because they hitchhike on clonal expansions that are driven by selectively advantageous alterations or by drift. In addition, data have confirmed that each cancer in each patient has an individually unique genomic profile. It is possible that cancer cells need only a modest number of phenotypic traits to deal with all of the constraints and evolve into a fully malignant or metastatic tumour[25], but the genomics data suggest that this can be achieved by an almost infinite variety of evolutionary trajectories and with multiple different combinations of driver mutations[44].

Paradoxically, genome profiles underestimate complexity. So far, they have been mostly one-off snapshots from a single sample at a single diagnostic time point. We know that serial or parallel sampling using more conventional genetic analysis uncovers genetic diversity within a tumour. Whole-genome sequencing of paired primary tumours versus metastatic samples has so far been limited, but it has revealed that individual metastatic lesions are clonal in origin and genetically unique, yet have a clonal ancestry traceable to the primary tumour[2]. 'The genome' description is perhaps also misleading because genetic variants are identified in 5–50% of reads, which suggests subclonal distribution of most mutations[53], but the segregation pattern of mutations within subclones is lost when DNA is extracted from the total cell population. This is important if patient-specific genomic profiles are to provide a platform for selecting therapeutic targets. Arguably, subclonal genetic diversity is key to the success or failure of therapy. This is a considerable challenge, technically and bioinformatically, in cancer genomics and will require deep sequencing[40]

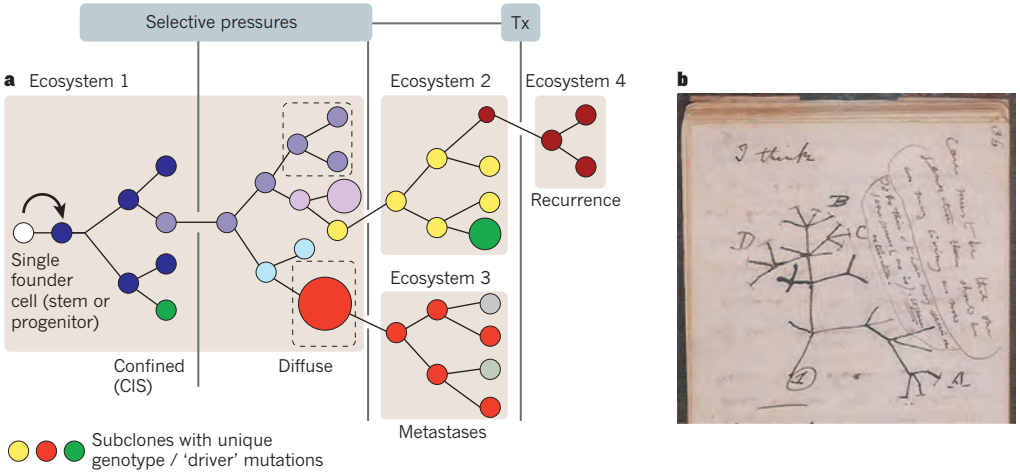

Figure 2 | The branching architecture of evolution. **a**, Cancer clones. Selective pressures allow some mutant subclones to expand while others become extinct or remain dormant. Vertical lines represents restraints or selective pressures. This is a representative pattern for common, solid cancers; as recognized by Nowell[3], leukaemic clones may expand over a shorter time frame (years versus decades), and be subject to fewer restraints and mutational events. Ecosystems 1–4 (boxes) represent the different tissue ecosystems or habitats. Smaller boxes within Ecosystem 1 represent localized habitats or niches. Each differently coloured circle represents a genetically distinct subclone. Metastatic subclones can branch off into different time points in the sequence from either minor or major clones in the primary tumour. Tx, therapy. CIS, carcinoma *in situ*. **b**, Darwin's branching evolutionary tree of speciation from his 1837 notebook.

and investigation of the genomes of single cells for patterns of segregation of mutations to understand the genetic diversity within neoplasms and how this changes in response to interventions.

Subclonal segregation of mutations and clonal architecture

The classic model of clonal evolution suggests there is a sequential acquisition of mutations with concomitant, successive subclonal dominance or selective sweeps. Histopathological evidence of disease progression (adenoma, carcinoma and metastases) supports this model. At each stage of this evolution, individual cells and their progeny (subclones) compete for space and resources. Multiplexed, single-cell mutational analysis (ideally in serial samples) is the most appropriate way to examine clonal architecture. So far, there are only a few examples of this[10,32,33], but they have provided evidence of the complex pattern of subclonal segregation of mutations — consistent with Nowell's model. The large amount of data from tissue sections, small biopsies and, more recently, single-cell analysis[33] is evidence that the evolutionary trajectories are complex and branching, exactly as Nowell proposed and in parallel with Darwin's iconic evolutionary speciation tree (Fig. 2). Attempts to simplify this complex system into a linear sequence of mutational events on the basis of cross-sectional data have probably been misleading[54]. However, by comparing the mutational genomes of the subclones, it is possible to discover their evolutionary or ancestral relationships, as well as the order of events during the development of that neoplasm[32,33,37,53,54]. Clonal evolution from common ancestral cancer cells is demonstrated in identical twins with concordant acute leukaemia[55,56], in metastatic lesions[2,10] and, by inference, in some cases of bilateral testicular cancer[57] (Fig. 3). In this context, divergent cancer-clone genotypes and phenotypes correspond to allopatric speciation in separate natural habitats (for example, Darwin's finches on the Galapagos Islands[58]).

Profiles of subclones within a neoplasm can be used to determine 'molecular clocks' that can then be linked to time events in the history of the neoplasm. For example, DNA methylation changes and base-pair mutations have been used to infer clonal expansion dynamics[26] and the time between initiation, invasion and metastasis[17,37,52]. It is even possible to determine the relative timing of events during progression from a single sample, based on deep sequencing[59].

Subclones may be mixed together within the primary tissue[37,60], but given their single-cell origin and bifurcating pathways, it is not surprising that they can also occupy distinctive territories[35,37,61,62] (Fig. 4a). Cancer-clone evolution involves contemporaneous subclones with distinctive mutational and phenotypic profiles that may be territorially segregated, which has considerable practical implications for diagnosis, prognosis and targeted therapy based on biopsy sampling[63]. It remains unclear whether all subclonal diversification reflects the impact of driver mutations and selective advantage, or is also the result of genetic drift of selectively neutral mutations or even epigenetic alterations. The level of diversity within the subclonal structure can be measured[35,64,65] and has been shown to be a robust biomarker for predicting progression to malignancy in Barrett's oesophagus[65]. It is also associated with the tumour stage and subtype of breast cancer[64].

Units of selection and cancer stem cells

Evolutionary theory suggests that natural selection operates in any system that has components with varying reproductive potential[4]. In

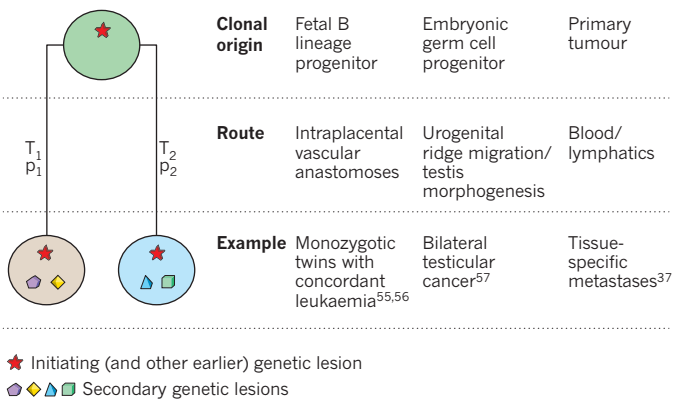

Figure 3 | Divergent (branching) clonal evolution of cancer with topographical separation. In each example, a clonal (single cell) ancestry is indicated by a shared acquired mutation (for example, *ETV6–RUNX1* fusion for leukaemias and *KIT* mutation for testicular cancers). The time at which the two subclones evolve (T_1 and T_2) can be temporarily synchronous or develop several years apart[37,55–57]. The probabilities of subclones emerging as shown are independent and different (p_1 and p_2). In most cases (90% for monozygotic twins), only one twin develops overt leukaemia. The penetrance of bilateral testicular cancer having a common origin[57] is unknown.

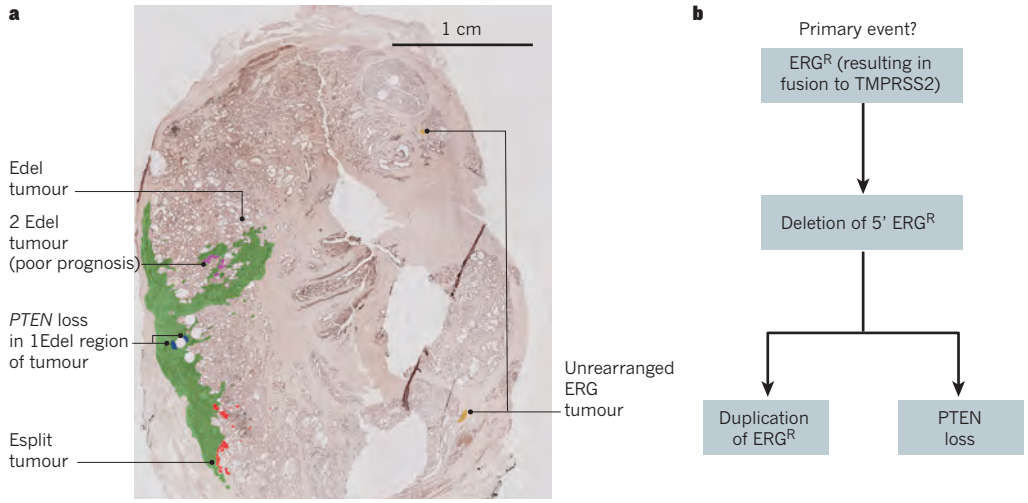

Figure 4 | Topography of cancer subclones. a, Tissue section of prostate to detect genetic events: *TMPRSS2–ERG* fusion (*ERG* via rearrangement (ERG^R)) and *PTEN* loss. **b**, The presumed sequence of clonal events.

the progression of cancer, or its resurgence after therapy, the primary unit of selection is the cell. This cell has to have extensive replicative potential, the so-called cancer stem cell (also known as the cancer-initiating or propagating cell) (Fig. 5).

The cancer stem-cell hypothesis was developed through transplantation experiments with leukaemic cells[66], and although it has been reported to be a general feature of all cancers[67], this idea is contentious. There has been no consensus on whether cancer stem cells are rare or high-frequency cells, or whether they have fixed, hierarchical or variable phenotypes, but considering the evolutionary progression in cancer, cells with extensive propagating activity are unlikely to be fixed entities[68,69]. Cancer stem cells are the cellular drivers of subclonal expansion and so probably vary in frequency and phenotypic features. The only feature they must have is the potential for extensive self-renewal (Fig. 5). Quantitative measures of stem-cell activity or self-renewal (through xenotransplantation or gene-expression signatures) can be used to predict the clinical outcome of several cancer types[70]. The cancer stem cell's ability to self-renew is made stronger by an aberrant genotype and, possibly, other, epigenetic, features. Several testable predictions can be made from this. First, cancer stem cells should evolve and change in genotype and phenotype as the cancer evolves before and after therapy. Some therapies may even provide a strong selection for cancer stem-cell survival and proliferation[71]. Second, as cancers progress, there should be selective pressure for the cells with the most extensive self-renewing capacity, but at the expense of cells with the ability to differentiate. This has been observed in chronic myeloid leukaemia (CML)[72] and mouse models[73,74]. A higher probability of symmetrical self-renewing proliferative cycles would be expected to result in an increased number and frequency of cancer stem cells. It is therefore of some consequence that loss of the *TP53* DNA damage checkpoint, which frequently correlates with cancer progression and clinical intransigence[75], seems to 'release' stem-cell-like transcriptional signatures[76] and leads to enhanced self-renewal in mammosphere culture systems[77]. The frequency of cancer stem cells could then increase from low to very high frequency as the disease progresses[78,79]. Third, for selection to operate through microenvironmental or therapeutic pressures, there should be contemporaneous genetic variation in cancer stem cells, which has been shown in leukaemias[33,80].

These considerations have significant clinical implications. Whatever the frequency and phenotype, if self-renewing cancer stem cells drive and sustain cancer-clone evolution, this suggests they are the repository of functionally relevant mutational events that drive clonal selection before and after therapy. This supports the view that cancer stem-cell restraint or elimination should be the aim of any therapy. However, if cancer stem cells are as genetically (and epigenetically) diverse as evolutionary considerations and initial experiments[33,80,81] indicate, this could be the reason for therapeutic failure. The adaptability of cancer stem cells provided by genetic diversity is added to by what seems to be their intrinsically lowered susceptibility to drugs and irradiation[82]. This may be because of the association with stromal cells[83] and the quiescence of cancer stem-cell subpopulations, as well as the properties of enhanced DNA repair and elevated expression of drug efflux pumps, which may be the evolved contingencies to protect normal stem cells.

Subclonal genetic heterogeneity is a common, if not universal, feature of cancers[84]. However, it cannot be assumed that all subclones are sustained by cancer stem cells; some could be evolutionary dead-ends generated by cells with only limited propagating potential. It is partly to accommodate this that the *in vivo* assay for cancer stem cells involves sequential transplants[66]. Ideally, the genomes of single cancer stem cells would be interrogated to investigate how they relate to subclones, but this is not currently possible. However, the genetic heterogeneity of cancer stem cells can be inferred by comparing subclonal diversity or clonal architecture before and after transplantation. Quadrant sections of glioblastoma have been shown to have divergent but related genotypes, but all sections contained cells that read-out in the *in vivo* (intracerebral) cancer stem-cell assay[85]. More definitive data come from comparing pre- and post-transplant subclonal genetic profiles that were investigated at the single cell level or by single nucleotide polymorphism arrays in B-cell precursor acute lymphoblastic leukaemia. Multiple subclones from each patient's diagnostic sample registered in the *in vivo* cancer stem-cell transplant assays, albeit with variable competitive potency[33,80,81]. We are still awaiting experimental confirmation that genetic diversity of cancer stem cells is a common feature of cancer, but, assuming that it is, this will have important therapeutic implications.

A Darwinian bypass

Nowell[3] stated in his landmark article "more research should be directed towards understanding and controlling the evolutionary process in tumours before it reaches the late stage seen in clinical cancer". Although cancer therapy has had its successes, in reality very few advanced or metastatic malignancies can be effectively controlled or eradicated. Genetic variation in cancer stem cells, particularly if induced by genetic instability, provides the opportunity for cells to escape and the therapy to fail. Other, non-genetic, mechanisms of positive selection by therapy exist, including signalling plasticity (or oncogene bypass)[86], quiescence[87] and epigenetic changes[88]; however, many of these depend on heritable, and thus selectable, epigenetic variation. Great expectation has been placed on the audit of cancer genomes that, by identifying recurrent and "druggable" mutations, would herald a new phase of highly specific or targeted

small-molecule inhibitors and personalized medicine[89]. Oncogene addiction may be the Achilles heel of cancer in this respect[90]. The success of imatinib and the derivative non-receptor tyrosine (ABL1) kinase inhibitors in CML[90] was very encouraging, but CML is not a typical cancer. It is essentially a premalignant (albeit ultimately lethal) condition, probably driven by a single founder mutation (*BCR–ABL1* fusion), which provides a universal target for therapy. Even in the most favourable of circumstances, escape occurs either by quiescence (and coupled resistance) of cancer stem cells[91] or by mutation of the ABL1 kinase target. Once CML has evolved to an overt malignancy or blast crisis, with increased genetic complexity, ABL1 kinase-directed therapy is often ineffective.

Other small-molecule inhibitors directed at mutant products have produced encouraging results in patients with advanced disease, but the benefits are transitory and cancer clones re-emerge with resistant features. When the targets selected are non-founder mutations, even if they are dominant in the neoplasm, therapy can be predicted to select for subclones lacking the mutant target[70]. Alternatively, subclones can have additional mutations that allow a bypass of the signalling pathway of the drug target, such as the MET proto-oncogene (*MET*) amplification in *EGFR* mutant lung cancer treated with EGFR kinase inhibitors[92].

Supporters of targeted therapy and personalized medicine argue that a combination of drugs that target components of networked signalling and are tailored to the individual patient's cancer genome is the solution to this problem. In this regard, synthetic lethal strategies seem promising[93].

Self-renewing cancer cells are the ultimate target for therapy, so high-throughput screening for selective inhibitors is an encouraging development[71]. Ways to target the components of the self-renewing process itself (independent of specific mutant genotype) deserve exploration, especially if a distinction can be made from normal adult stem cells. In the case of CML, intrinsically resistant (and possibly quiescent) stem cells, have been targeted by combining selective kinase (ABL1) inhibitors with inhibitors of a histone deacetylase[94] or BCL6 (ref. 95). Ultimately, it may prove difficult to thwart the plasticity and adaptability of cancer cells (or cancer stem cells), which are an inherent evolutionary feature of advanced disease, and a 'Darwinian bypass' may be required, for which there are a number of possibilities. An implication of the evolutionary diversity of cancer is that prevention (smoking cessation, avoiding sunburn, prophylactic vaccines, and so on) makes a great deal of sense, as does early detection and intervention (that is, before genetic diversification and dissemination become extensive).

An alternative therapeutic strategy is to focus on the microenvironmental habitat using 'ecological therapy', which aims to change the essential habitat and dependency of the cancer cells[96]. For example, anti-angiogenesis can provide a potent restraint on cancer stem cells[97]. Other examples are the use of bisphosphonates to remodel bone in patients with prostate cancer, the use of aromatase inhibitors in patients with breast cancer, exploiting hypoxia, the use of inhibitors of inflammation or tumour-infiltrating macrophages, and blocking cancer stem-cell interactions with essential stromal or niche components[96,98].

Another alternative is to control the cancer, rather than eradicate it, thereby turning cancer into a chronic disease. Because the speed of evolution is proportional to the fitness differential between the cells, cytotoxic drugs are predicted to select rapidly for resistance[5]. It is thought they cause competitive release[99] by removing all of the competitors of resistant cells. In contrast, cytostatic drugs should delay progression and mortality longer than cytotoxic drugs because sensitive competitor cells remain in the tissue to occupy space and consume resources that would otherwise be used by the resistant clones. In addition, by suppressing cell division, cytostatic drugs also suppress the opportunities for new mutations. A study by Gatenby and colleagues[100] showed that by treating an aggressive ovarian cancer (OVCAR-3) xenograft tumour to maintain a stable size, rather than to eradicate it, host mice could be kept alive much longer. Moreover, the dose of carboplatin necessary to keep the tumour at a manageable size declined over time[100]. Researchers should now focus on what phenotypes can be selected for to make neoplasms less deadly and more clinically manageable.

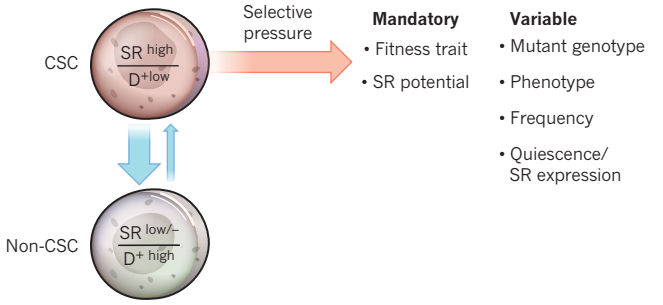

Figure 5 | Selective pressure on cancer stem cells. Selective pressures can include environmentally derived genotoxicity, natural or physiological restraints, cancer therapy, and so on. Mutation in progenitor cells can convert these cells back to a self-renewing population[72]. The small blue arrow represents a mutation; the large blue arrow represents differentiation: in both cases they represent a change in state. In addition to the mandatory trait of self-renewal, cancer stem cells (CSC), can exhibit any phenotypic feature that allows cells to continue to survive and proliferate in the face of a particular constraint. D^+, differentiation; SR, self-renewal.

The evolutionary theory of cancer has survived 35 years of empirical observation and testing, so today it could be considered a bona fide scientific theory. The basic components of somatic evolution are well understood, but the dynamics of somatic evolution remain unclear. Fortunately, there are evolutionary biology tools that may be applied to neoplasms to address many of the fundamental cancer biology questions, such as the order of events in progression, distinguishing driver from passenger mutations, and understanding and preventing therapeutic resistance. The dynamics of clonal diversification and selection are critical to understanding these issues. The challenge now is to use the clinical opportunities to address directly the evolutionary adaptability of neoplasms and design interventions to slow, direct or control cancer-cell evolution to delay or prevent mortality. ■

1. Jemal, A. *et al.* Cancer statistics, 2008. *CA Cancer J. Clin.* **58,** 71–96 (2008).
2. Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. *Science* **331,** 1553–1558 (2011).
3. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
The foundation paper that established the evolutionary theory of cancer.
4. Merlo, L. M., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nature Rev. Cancer* **6,** 924–935 (2006).
5. Pepper, J., Scott Findlay, C., Kassen, R., Spencer, S. & Maley, C. Cancer research meets evolutionary biology. *Evol. Appl.* **2,** 62–70 (2009).
6. Greaves, M. *Cancer: The Evolutionary Legacy* (Oxford Univ. Press, 2000).
7. Sakr, W. A., Haas, G. P., Cassin, B. F., Pontes, J. E. & Crissman, J. D. The frequency of carcinoma and intraepithelial neoplasia of the prostate in young male patients. *J. Urol.* **150,** 379–385 (1993).
8. Mori, H. *et al.* Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc. Natl Acad. Sci. USA* **99,** 8242–8247 (2002).
9. Reid, B. J., Li, X., Galipeau, P. C. & Vaughan, T. L. Barrett's oesophagus and oesophageal adenocarcinoma: time for a new synthesis. *Nature Rev. Cancer* **10,** 87–101 (2010).
10. Klein, C. A. Parallel progression of primary tumours and metastases. *Nature Rev. Cancer* **9,** 302–312 (2009).
11. Malaise, E. P., Chavaudra, N. & Tubiana, M. The relationship between growth rate, labelling index and histological type of human solid tumours. *Eur. J. Cancer* **9,** 305–312 (1973).
12. Tsai, A. G. *et al.* Human chromosomal translocations at CpG sites and a theoretical basis for their lineage and stage specificity. *Cell* **135,** 1130–1142 (2008).
13. Bardelli, A. *et al.* Carcinogen-specific induction of genetic instability. *Proc. Natl Acad. Sci. USA* **98,** 5770–5775 (2001).
14. Cahill, D. P., Kinzler, K. W., Vogelstein, B. & Lengauer, C. Genetic instability and Darwinian selection in tumors. *Trends Cell Biol.* **9,** M57–M60 (1999).

15. Barcellos-Hoff, M. H., Park, C. & Wright, E. G. Radiation and the microenvironment – tumorigenesis and therapy. *Nature Rev. Cancer* **5,** 867–875 (2005).
16. Maley, C. C. *et al.* Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's esophagus. *Cancer Res.* **64,** 3414–3427 (2004).
17. Tao, Y. *et al.* Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data. *Proc. Natl Acad. Sci. USA* **108,** 12042–12047 (2011).
18. Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463,** 893–898 (2010).
19. Youn, A. & Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27,** 175–181 (2011).
20. Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* **173,** 2187–2198 (2006).
21. Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl Acad. Sci. USA* **107,** 18545–18550 (2010).
22. Schwartz, M., Zlotorynski, E. & Kerem, B. The molecular basis of common and rare fragile sites. *Cancer Lett.* **232,** 13–26 (2006).
23. Loeb, L. A. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nature Rev. Cancer* **11,** 450–457 (2011).
24. Weisenberger, D. J. *et al.* CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with *BRAF* mutation in colorectal cancer. *Nature Genet* **38,** 787–793 (2006).
25. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
This paper consolidates the common phenotypes that evolve in neoplastic cells of all types.
26. Siegmund, K. D., Marjoram, P., Woo, Y. J., Tavare, S. & Shibata, D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc. Natl Acad. Sci. USA* **106,** 4828–4833 (2009).
27. Varley, K. E., Mutch, D. G., Edmonston, T. B., Goodfellow, P. J. & Mitra, R. D. Intra-tumor heterogeneity of *MLH1* promoter methylation revealed by deep single molecule bisulfite sequencing. *Nucleic Acids Res.* **37,** 4603–4612 (2009).
28. Aktipis, C. A., Kwan, V. S. Y., Johnson, K. A., Neuberg, S. L. & Maley, C. C. Overlooking evolution: a systematic analysis of cancer relapse and therapeutic resistance research. *PLoS ONE* **6,** e261000 (2011).
29. Beerenwinkel, N. *et al.* Genetic progression and the waiting time to cancer. *PLoS Comput. Biol.* **3,** e225 (2007).
30. de Visser, J. A. & Rozen, D. E. Clonal interference and the periodic selection of new beneficial mutations in *Escherichia coli*. *Genetics* **172,** 2093–2100 (2006).
31. Leedham, S. J. *et al.* Individual crypt genetic heterogeneity and the origin of metaplastic glandular epithelium in human Barrett's oesophagus. *Gut* **57,** 1041–1048 (2008).
32. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–94 (2011).
Single-cell sequencing revealed the clonal structure of two breast cancers.
33. Anderson, K. *et al.* Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469,** 356–361 (2011).
Single-cell genetic analyses and xenografts revealed the clonal architecture within acute lymphoblastic leukaemia stem-cell populations and demonstrated repeated independent acquisition of copy number changes within the same neoplasm.
34. Tsao, J. L. *et al.* Colorectal adenoma and cancer divergence. Evidence of multilineage progression. *Am. J. Pathol.* **154,** 1815–1824 (1999).
35. Maley, C. C. *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature Genet.* **38,** 468–473 (2006).
36. Sidransky, D. *et al.* Clonal expansion of p53 mutant cells is associated with brain tumour progression. *Nature* **355,** 846–847 (1992).
37. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467,** 1114–1117 (2010).
38. Gould, S. J. & Eldredge, N. Punctuated equilibrium comes of age. *Nature* **366,** 223–227 (1993).
39. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144,** 27–40 (2011).
40. Campbell, P. J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl Acad. Sci. USA* **105,** 13081–13086 (2008).
Deep sequencing revealed rare (frequency <0.001) intermediate genotypes between the common clones in leukaemias (using immunoglobulin rearrangements as surrogate mutations).
41. Aguirre-Ghiso, J. A. Models, mechanisms and clinical evidence for cancer dormancy. *Nature Rev. Cancer* **7,** 834–846 (2007).
42. Isoda, T. *et al.* Immunologically silent cancer clone transmission from mother to offspring. *Proc. Natl Acad. Sci. USA* **106,** 17882–17885 (2009).
43. Welsh, J. S. Contagious cancer. *Oncologist* **16,** 1–4 (2011).
44. Gatenby, R. A. & Gillies, R. J. A microenvironmental model of carcinogenesis. *Nature Rev. Cancer* **8,** 56–61 (2008).
45. Bierie, B. & Moses, H. L. Tumour microenvironment: TGFβ: the molecular Jekyll and Hyde of cancer. *Nature Rev. Cancer* **6,** 506–520 (2006).
46. Lathia, J. D., Heddleston, J. M., Venere, M. & Rich, J. N. Deadly teamwork: neural cancer stem cells and the tumor microenvironment. *Cell Stem Cell* **8,** 482–485 (2011).
47. Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255,** 197–200 (1975).
This paper identified natural selection as a driving force in carcinogenesis and identified tissue architecture as a cancer suppressor, and posited an immortal strand of DNA in tissue stem cells.
48. Anderson, A. R., Weaver, A. M., Cummings, P. T. & Quaranta, V. Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment. *Cell* **127,** 905–915 (2006).
49. Chen, J., Sprouffske, K., Huang, Q. & Maley, C. C. Solving the puzzle of metastasis: the evolution of cell migration in neoplasms. *PLoS ONE* **6,** e17933 (2011).
50. Mazzone, M. *et al.* Heterozygous deficiency of *PHD2* restores tumor oxygenation and inhibits metastasis via endothelial normalization. *Cell* **136,** 839–851 (2009).
51. Gilbert, L. A. & Hemann, M. T. DNA damage-mediated induction of a chemoresistant niche. *Cell* **143,** 355–366 (2010).
52. Jones, S. *et al.* Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl Acad. Sci. USA* **105,** 4283–4288 (2008).
53. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).
54. Sprouffske, K., Pepper, J. W. & Maley, C. C. Accurate reconstruction of the temporal order of mutations in neoplastic progression. *Cancer Prev. Res.* **4,** 1135–1144 (2011).
55. Greaves, M. F., Maia, A. T., Wiemels, J. L. & Ford, A. M. Leukemia in twins: lessons in natural history. *Blood* **102,** 2321–2333 (2003).
56. Bateman, C. M. *et al.* Acquisition of genome-wide copy number alterations in monozygotic twins with acute lymphoblastic leukemia. *Blood* **115,** 3553–3558 (2010).
57. Oosterhuis, J. W. & Looijenga, L. H. Testicular germ-cell tumours in a broader perspective. *Nature Rev. Cancer* **5,** 210–222 (2005).
58. Grant, P. R. & Grant, B. R. *How and Why Species Multiply* (Princeton Univ. Press, 2008).
59. Durinck, S. *et al.* Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov.* **1,** 137–143 (2011).
60. Gonzalez-Garcia, I., Sole, R. V. & Costa, J. Metapopulation dynamics and spatial heterogeneity in cancer. *Proc. Natl Acad. Sci. USA* **99,** 13085–13089 (2002).
61. Clark, J. *et al.* Complex patterns of *ETS* gene alteration arise during cancer development in the human prostate. *Oncogene* **27,** 1993–2003 (2008).
62. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20,** 68–80 (2010).
63. Allred, D. C. *et al.* Ductal carcinoma *in situ* and the emergence of diversity during breast cancer evolution. *Clin. Cancer Res.* **14,** 370–378 (2008).
64. Park, S. Y., Gonen, M., Kim, H. J., Michor, F. & Polyak, K. Cellular and genetic diversity in the progression of *in situ* human breast carcinomas to an invasive phenotype. *J. Clin. Invest.* **120,** 636–644 (2010).
65. Merlo, L. M. *et al.* A comprehensive survey of clonal diversity measures in Barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. *Cancer Prev. Res.* **3,** 1388–1397 (2010).
66. Dick, J. E. Stem cell concepts renew cancer research. *Blood* **112,** 4793–4807 (2008).
67. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414,** 105–111 (2001).
68. Greaves, M. Cancer stem cells renew their impact. *Nature Med.* **17,** 1046–1048 (2011).
69. Rosen, J. M. & Jordan, C. T. The increasing complexity of the cancer stem cell paradigm. *Science* **324,** 1670–1673 (2009).
70. Greaves, M. Cancer stem cells: back to Darwin? *Semin. Cancer Biol.* **20,** 65–70 (2010).
71. Gupta, P. B. *et al.* Identification of selective inhibitors of cancer stem cells by high-throughput screening. *Cell* **138,** 645–659 (2009).
72. Jamieson, C. H. *et al.* Granulocyte–macrophage progenitors as candidate leukemic stem cells in blast-crisis CML. *N. Engl. J. Med.* **351,** 657–667 (2004).
73. Akala, O. O. *et al.* Long-term haematopoietic reconstitution by *Trp53*$^{-/-}$ *p16*$^{Ink4a-/-}$*p19*$^{Arf-/-}$ multipotent progenitors. *Nature* **453,** 228–232 (2008).
74. Krivtsov, A. V. *et al.* Transformation from committed progenitor to leukaemia stem cell initiated by MLL–AF9. *Nature* **442,** 818–822 (2006).
75. Olivier, M. & Taniere, P. Somatic mutations in cancer prognosis and prediction: lessons from *TP53* and *EGFR* genes. *Curr. Opin. Oncol.* **23,** 88–92 (2011).
76. Mizuno, H., Spike, B. T., Wahl, G. M. & Levine, A. J. Inactivation of *p53* in breast cancers correlates with stem cell transcriptional signatures. *Proc. Natl Acad. Sci. USA* **107,** 22745–22750 (2010).
77. Cicalese, A. *et al.* The tumor suppressor *p53* regulates polarity of self-renewing divisions in mammary stem cells. *Cell* **138,** 1083–1095 (2009).
78. Quintana, E. *et al.* Efficient tumour formation by single human melanoma cells. *Nature* **456,** 593–598 (2008).
New xenograft methods revealed that cancer stem cells are common cell types in melanoma.
79. Pece, S. *et al.* Biological and molecular heterogeneity of breast cancers correlates with their cancer stem cell content. *Cell* **140,** 62–73 (2010).
80. Notta, F. *et al.* Evolution of human *BCR–ABL1* lymphoblastic leukaemia-initiating cells. *Nature* **469,** 362–367 (2011).
81. Clappier, E. *et al.* Clonal selection in xenografted human T cell acute

lymphoblastic leukemia recapitulates gain of malignancy at relapse. *J. Exp. Med.* **208,** 653–661 (2011).
82. Frank, N. Y., Schatton, T. & Frank, M. H. The therapeutic promise of the cancer stem cell concept. *J. Clin. Invest.* **120,** 41–50 (2010).
83. Ishikawa, F. *et al.* Chemotherapy-resistant human AML stem cells home to and engraft within the bone-marrow endosteal region. *Nature Biotechnol.* **25,** 1315–1321 (2007).
84. Marusyk, A. & Polyak, K. Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta* **1805,** 105–117 (2010).
85. Piccirillo, S. G. M. *et al.* Distinct pools of cancer stem-like cells coexist within human glioblastomas and display different tumorigenicity and independent genomic evolution. *Oncogene* **28,** 1807–1811 (2009).
86. Solit, D. & Sawyers, C. L. How melanomas bypass new therapy. *Nature* **468,** 902–903 (2010).
87. Goff, D. & Jamieson, C. Cycling toward elimination of leukemic stem cells. *Cell Stem Cell* **6,** 296–297 (2010).
88. Sharma, S. V. *et al.* A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* **141,** 69–80 (2010).
89. Chin, L., Andersen, J. N. & Futreal, P. A. Cancer genomics: from discovery science to personalized medicine. *Nature Med.* **17,** 297–303 (2011).
90. Sawyers, C. L. Shifting paradigms: the seeds of oncogene addiction. *Nature Med.* **15,** 1158–1161 (2009).
91. Graham, S. M. *et al.* Primitive, quiescent, Philadelphia-positive stem cells from patients with chronic myeloid leukemia are insensitive to STI571 *in vitro*. *Blood* **99,** 319–325 (2002).
92. Turke, A. B. *et al.* Preexistence and clonal selection of *MET* amplification in *EGFR* mutant NSCLC. *Cancer Cell* **17,** 77–88 (2010).
93. Ashworth, A., Lord, C. J. & Reis-Filho, J. S. Genetic interactions in cancer progression and treatment. *Cell* **145,** 30–38 (2011).
94. Zhang, B. *et al.* Effective targeting of quiescent chronic myelogenous leukemia stem cells by histone deacetylase inhibitors in combination with imatinib mesylate. *Cancer Cell* **17,** 427–442 (2010).
95. Duy, C. *et al. BCL6* enables Ph^+ acute lymphoblastic leukaemia cells to survive *BCR–ABL1* kinase inhibition. *Nature* **473,** 384–388 (2011).
96. Pienta, K. J., McGregor, N., Axelrod, R. & Axelrod, D. E. Ecological therapy for cancer: defining tumors using an ecosystem paradigm suggests new opportunities for novel cancer treatments. *Trans. Oncol.* **1,** 158–164 (2008).
97. Calabrese, C. *et al.* A perivascular niche for brain tumor stem cells. *Cancer Cell* **11,** 69–82 (2007).
98. Bissell, M. J. & Hines, W. C. Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression. *Nature Med.* **17,** 320–329 (2011).
99. Wargo, A. R., Huijben, S., de Roode, J. C., Shepherd, J. & Read, A. F. Competitive release and facilitation of drug-resistant parasites after therapeutic chemotherapy in a rodent malaria model. *Proc. Natl Acad. Sci. USA* **104,** 19914–19919 (2007).
100. Gatenby, R. A., Silva, A. S., Gillies, R. J. & Frieden, B. R. Adaptive therapy. *Cancer Res.* **69,** 4894–4903 (2009).
Dosing to maintain tumour size prolonged survival far longer than high-dose therapy in a mouse xenograft model.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The research of M.G. is supported by Leukaemia & Lymphoma Research UK and The Kay Kendall Leukaemia Fund. The research of C.M. is supported by Research Scholar Grant #117209-RSG-09-163-01-CNE from the American Cancer Society and NIH grants P01 CA91955, U54 CA143803, R01 CA149566 and R01 CA140657. The authors thank C.Cooper and J.Clark for the use of the image in Fig. 4.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to M.G. (mel.greaves@icr.ac.uk).

PLOS | BIOLOGY

PERSPECTIVE

Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

1 Coordinated Science Laboratory and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **2** Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **3** Carl R. Woese Institute for Genomic Biology & Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **4** School of Library and Information Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **5** Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, **6** Carl R. Woese Institute for Genomic Biology, Department of Entomology, and Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America

* mschatz@cshl.edu (MCS); sinhas@illinois.edu (SS); generobi@illinois.edu (GER)

Citation: Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195

Published: July 7, 2015

Funding: This research was supported, in part, by grant 1U54GM114838 awarded by NIGMS to SS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative and by National Institutes of Health award (R01-HG006677) to MCS. ZDS and RKI were supported by NSF grant MRI13-37732 (S.S. Lumetta, PI). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: API, application programming interface; ASKAP, Australian Square Kilometre Array Pathfinder; CPU, central processing unit; ExAC, Exome Aggregation Consortium; ICGC, International Cancer Genome Consortium; I/O, input/output; NIH/NCBI, National Institutes of Health National Center for Biotechnology Information; SKA, Square Kilometre Array; SRA, Sequence Read Archive; TCGA, The

Abstract

Genomics is a Big Data science and is going to get much bigger, very soon, but it is not known whether the needs of genomics will exceed other Big Data domains. Projecting to the year 2025, we compared genomics with three other major generators of Big Data: astronomy, YouTube, and Twitter. Our estimates show that genomics is a "four-headed beast"—it is either on par with or the most demanding of the domains analyzed here in terms of data acquisition, storage, distribution, and analysis. We discuss aspects of new technologies that will need to be developed to rise up and meet the computational challenges that genomics poses for the near future. Now is the time for concerted, community-wide planning for the "genomical" challenges of the next decade.

We compared genomics with three other major generators of Big Data: astronomy, YouTube, and Twitter. Astronomy has faced the challenges of Big Data for over 20 years and continues with ever-more ambitious studies of the universe. YouTube burst on the scene in 2005 and has sparked extraordinary worldwide interest in creating and sharing huge numbers of videos. Twitter, created in 2006, has become the poster child of the burgeoning movement in computational social science [6], with unprecedented opportunities for new insights by mining the enormous and ever-growing amount of textual data [7]. Particle physics also produces massive quantities of raw data, although the footprint is surprisingly limited since the vast majority of data are discarded soon after acquisition using the processing power that is coupled to the sensors [8]. Consequently, we do not include the domain in full detail here, although that model of rapid filtering and analysis will surely play an increasingly important role in genomics as the field matures.

Cancer Genome Atlas; WGS, whole genome shotgun.

To compare these four disparate domains, we considered the four components that comprise the "life cycle" of a dataset: acquisition, storage, distribution, and analysis (Table 1).

Data Acquisition

The four Big Data domains differ sharply in how data are acquired. Most astronomy data are acquired from a few highly centralized facilities [9]. By contrast, YouTube and Twitter acquire data in a highly distributed manner, but under a few standardized protocols. Astronomy, YouTube, and Twitter are expected to show continued dramatic growth in the volume of data to be acquired. For example, the Australian Square Kilometre Array Pathfinder (ASKAP) project currently acquires 7.5 terabytes/second of sample image data, a rate projected to increase 100-fold to 750 terabytes/second (~25 zettabytes per year) by 2025 [9,10]. YouTube currently has 300 hours of video being uploaded every minute, and this could grow to 1,000–1,700 hours per minute (1–2 exabytes of video data per year) by 2025 if we extrapolate from current trends (S1 Note). Today, Twitter generates 500 million tweets/day, each about 3 kilobytes including metadata (S2 Note). While this figure is beginning to plateau, a projected logarithmic growth rate would suggest a 2.4-fold growth by 2025, to 1.2 billion tweets per day, 1.36 petabytes/year. In short, data acquisition in these domains is expected to grow by up to two orders of magnitude in the next decade.

For genomics, data acquisition is highly distributed and involves heterogeneous formats. The rate of growth over the last decade has also been truly astonishing, with the total amount of sequence data produced doubling approximately every seven months (Fig 1). The Omics-Maps catalog of all known sequencing instruments in the world [11] reports that currently there are more than 2,500 high-throughput instruments, manufactured by several different companies, located in nearly 1,000 sequencing centers in 55 countries in universities, hospitals, and other research laboratories. These centers range in size from small laboratories with a few instruments generating a few terabases per year to large dedicated facilities producing several petabases a year. (An approximate conversion factor to use in interpreting these numbers is 4 bases = 1 byte, though we will revisit this below.)

The raw sequencing reads used in most published studies are archived at either the Sequence Read Archive (SRA) maintained by the United States National Institutes of Health National Center for Biotechnology Information (NIH/NCBI) or one of the international counterparts. The SRA currently contains more than 3.6 petabases of raw sequence data (S1 Fig), which reflects the ~32,000 microbial genomes, ~5,000 plant and animal genomes, and ~250,000 individual human genomes that have been sequenced or are in progress thus far [12].

Table 1. Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

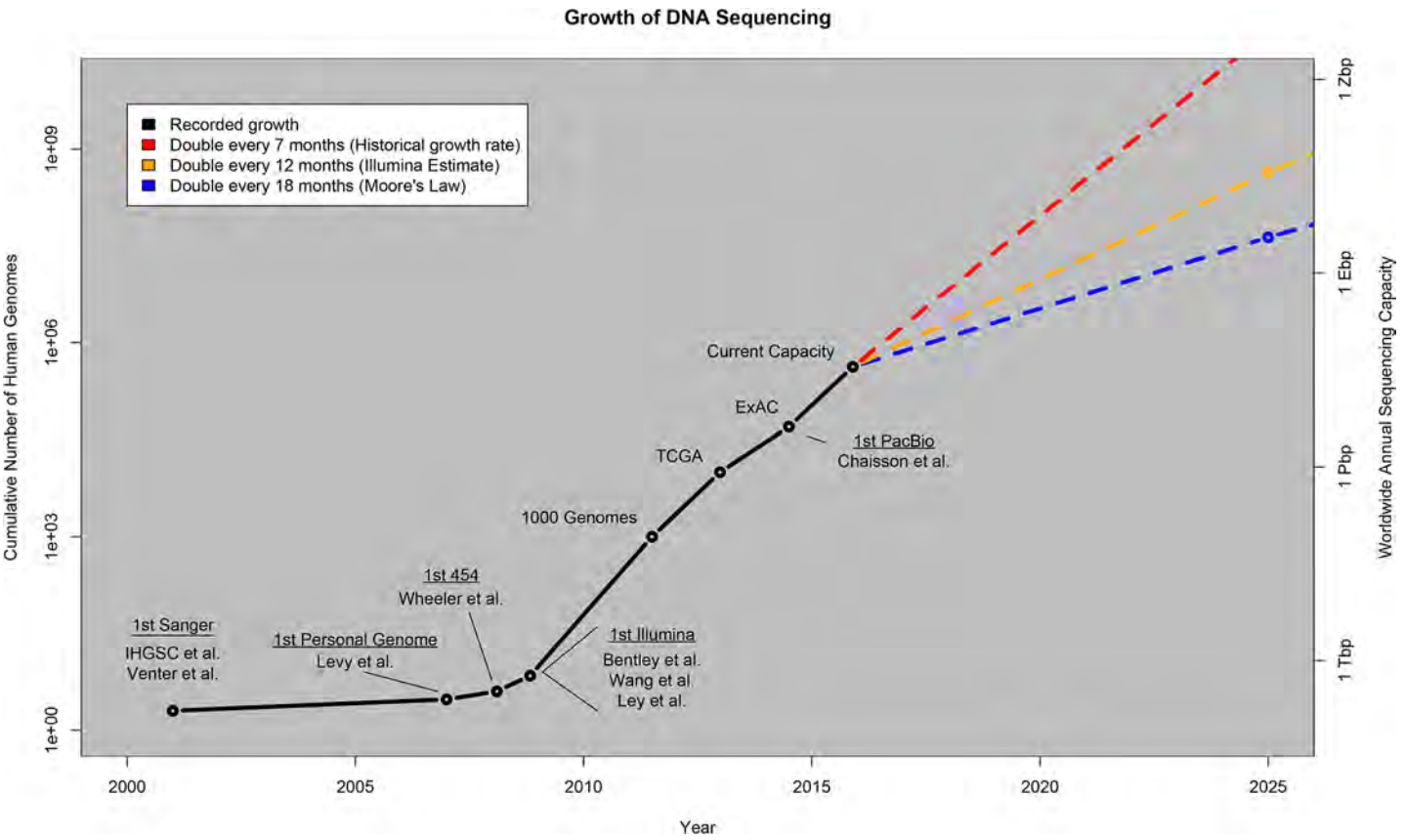

Fig 1. Growth of DNA sequencing. The plot shows the growth of DNA sequencing both in the total number of human genomes sequenced (left axis) as well as the worldwide annual sequencing capacity (right axis: Tera-basepairs (Tbp), Peta-basepairs (Pbp), Exa-basepairs (Ebp), Zetta-basepairs (Zbps)). The values through 2015 are based on the historical publication record, with selected milestones in sequencing (first Sanger through first PacBio human genome published) as well as three exemplar projects using large-scale sequencing: the 1000 Genomes Project, aggregating hundreds of human genomes by 2012 [3]; The Cancer Genome Atlas (TCGA), aggregating over several thousand tumor/normal genome pairs [4]; and the Exome Aggregation Consortium (ExAC), aggregating over 60,000 human exomes [5]. Many of the genomes sequenced to date have been whole exome rather than whole genome, but we expect the ratio to be increasingly favored towards whole genome in the future. The values beyond 2015 represent our projection under three possible growth curves as described in the main text.

doi:10.1371/journal.pbio.1002195.g001

However, the 3.6 petabases represent a small fraction of the total produced; most of it is not yet in these archives. Based on manufacturer specifications of the instruments, we estimate the current worldwide sequencing capacity to exceed 35 petabases per year, including the sixteen Illumina X-Ten systems that have been sold so far [13], each with a capacity of ~2 petabases per year [14].

Over the next ten years, we expect sequencing capacities will continue to grow very rapidly, although the project growth becomes more unpredictable the further out we consider. If the growth continues at the current rate by doubling every seven months, then we should reach more than one exabase of sequence per year in the next five years and approach one zettabase of sequence per year by 2025 (Fig 1, Table 1). Interestingly, even at the more conservative estimates of doubling every 12 months (Illumina's current own estimate [12]) or every 18 months (equivalent to Moore's law), we should reach exabase-scale genomics well within the next decade. We anticipate this sequencing will encompass genome sequences for most of the approximately 1.2 million described species of plants and animals [15]. With these genomes, plus those of thousands of individuals of "high value" species for energy, environmental, and agricultural reasons, we estimate that there will be at least 2.5 million plant and animal genome

sequences by 2025. For example, the genomics powerhouse BGI, in conjunction with the International Rice Research Institute and the Chinese Academy of Agricultural Sciences, has already sequenced 3,000 varieties of rice [16] and announced a massive project of their own to sequence one million plant and animal genomes [17]. The Smithsonian Institute also has similar plans to "capture and catalog all the DNA from the world's flora and fauna." There also will be genomes for several millions of microbes, with explosive growth projected for both medical and environmental microbe metagenomic sequencing [18,19].

These estimates, however, are dwarfed by the very reasonable possibility that a significant fraction of the world's human population will have their genomes sequenced. The leading driver of this trend is the promise of genomic medicine to revolutionize the diagnosis and treatment of disease, with some countries contemplating sequencing large portions of their populations: both England [20] and Saudi Arabia [21] have announced plans to sequence 100,000 of their citizens, one-third of Iceland's 320,000 citizens have donated blood for genetic testing [22], and researchers in both the US [23] and China [17] both aim to sequence 1 million genomes in the next few years. With the world's population projected to top 8 billion by 2025, it is possible that as many as 25% of the population in developed nations and half of that in less-developed nations will have their genomes sequenced (comparable to the current worldwide distribution of Internet users [24]).

We therefore estimate between 100 million and as many as 2 billion human genomes could be sequenced by 2025, representing four to five orders of magnitude growth in ten years and far exceeding the growth for the three other Big Data domains. Indeed, this number could grow even larger, especially since new single-cell genome sequencing technologies are starting to reveal previously unimagined levels of variation, especially in cancers, necessitating sequencing the genomes of thousands of separate cells in a single tumor [10].

Moreover, the technology used to sequence DNA is deployed creatively for other applications (e.g., transcriptome, epigenome, proteome, metabolome, and microbiome sequencing) necessitating generating new sequencing data multiple times per person to monitor molecular activity [25]. These applications require precise quantitative counts of sequencing reads to capture diversity of expression or diversity of abundances, thus requiring millions of reads to accurately estimate underlying distributions as they change over time. For medicine, just having the genome will not be sufficient: for each individual, it will need to be coupled with other relevant 'omics data sets, some collected periodically and from different tissues, to compare healthy and diseased states [26]. Computational challenges will increase because of dramatic increases in the total volume of genomic data per person, as will the complexities of integrating these diverse data sources to improve health and cure diseases. Genomics thus appears to pose the greatest challenges for data acquisition of the four Big Data domains.

Data Storage

Data storage requirements for all four domains are projected to be enormous. Today, the largest astronomy data center devotes ~100 petabytes to storage, and the completion of the Square Kilometre Array (SKA) project is expected to lead to a storage demand of 1 exabyte per year. YouTube currently requires from 100 petabytes to 1 exabyte for storage and may be projected to require between 1 and 2 exabytes additional storage per year by 2025. Twitter's storage needs today are estimated at 0.5 petabytes per year, which may increase to 1.5 petabytes in the next ten years. (Our estimates here ignore the "replication factor" that multiplies storage needs by ~4, for redundancy.) For genomics, we have determined more than 100 petabytes of storage are currently used by only 20 of the largest institutions (S1 Table).

Projections of storage requirements for sequence data depend on the accuracy and application of the sequencing. For every 3 billion bases of human genome sequence, 30-fold more data (~100 gigabases) must be collected because of errors in sequencing, base calling, and genome alignment. This means that as much as 2–40 exabytes of storage capacity will be needed by 2025 just for the human genomes (S3 Note). These needs can be diminished with effective data compression [27], but decompression times and fidelity are a major concern in compressive genomics [28].

Are the emerging "third-generation" single-molecule sequencing technologies with much longer reads, such as those from Pacific Biosciences and Oxford Nanopore, a computational panacea? Though error rates currently are higher and throughput lower than short-read technologies, as they mature, these technologies are starting to be used to sequence and assemble nearly entire chromosomes [29]. This will minimize the need to oversample as much, and eventually, the raw sequence data may not need to be stored at all. However, eliminating the need to store raw sequence data and only retaining complete genomes will have relatively little impact overall—perhaps one or two orders of magnitude less data storage. More significant reductions in storage demand will come when improvements in sequencing accuracy and database comprehensiveness reach the point at which genome sequences themselves do not need to be stored, just the list of variants relative to a reference collection ("delta encoding") [30]. This works well for cataloging the simplest variants in a human genome, but it may not be as useful for complex samples, such as cancer genomes, that have many novel rearrangements and mutations. While certainly helpful, we thus do not expect long-read sequencing technology or delta encoding to solve the storage challenges for genome sequencing in 2025.

In contrast, we do see great opportunities for data reduction and real-time analysis of other 'omics analysis. For example, once sequencing becomes fast enough and the methods mature enough to correctly infer transcript expression levels in real time, we anticipate that raw RNA-seq reads will no longer be stored, except for specific research purposes. Already several such "streaming" algorithms have been published for this purpose, performing as well as or superior to their nonstreaming counterparts [31]. For RNA-seq and other 'omics applications, genomics will benefit greatly from the lessons learned in particle physics, in which in most cases raw data are discarded almost as fast as they are generated in favor of higher level and greatly compressed summaries.

Altogether, we anticipate the development of huge genomics archives used for storing millions of genomes along with the associated 'omics measurements over time. Ideally, these archives will also collect or be linked to the patient phenotypic data, especially disease outcomes and treatments provided to support retrospective analysis as new relationships are discovered. To make it practical to search and query through such vast collections, the data will be stored in hierarchical systems that make data and their statistical summaries available at different levels of compression and latency, as used in astronomy [32] and text analysis [33]. Thus, although total genomic data could far exceed the demands for the others, with the right new innovations the net requirements could be similar to the domains of astronomy and YouTube.

Data Distribution

Astronomy, YouTube, Twitter, and genomics also differ greatly in data distribution patterns. The major bandwidth requirement of the SKA project is to get data from its 3,000 antennae to a central server, requiring as much as 600 terabytes/second [34]. The bandwidth usage of YouTube is relatively small for a single download and well supported by the average consumer's 10 Mbps connection, but aggregate needs worldwide are enormous, with estimates up to 240

petabytes/day (S4 Note) [35]. The distribution patterns of genomics data are much more heterogeneous, involving elements of both situations [36].

Genomic data are distributed in units spanning a wide range of sizes, from comparisons of a few bases or gene sequences to large multiterabyte bulk downloads from central repositories. For large-scale analysis, cloud computing is particularly suited to decreasing the bandwidth for distribution of genomic data [37] so that applications can run on remote machines that already have data [38]. Only small segments of code are uploaded and highly processed outputs are downloaded, thus significantly reducing the computing resources necessary for distribution.

But in addition to tailoring genomics applications for the cloud, new methods of data reliability and security are required to ensure privacy, much more so than for the other three domains. A serious breach of medically sensitive genomic data would have permanent consequences and could seriously hinder the development of genomic medicine. Homomorphic encryption systems, in which encrypted data can be analyzed and manipulated for certain controlled queries without disclosing the raw data, are currently too computationally expensive for widespread use, but these and related cryptographic techniques are promising areas of research [39].

Data Analysis

Astronomy, YouTube, Twitter, and genomics differ most in computational requirements for data analysis. Astronomy data require extensive specialized analysis, but the bulk of this requirement is for in situ processing and reduction of data by computers located near the telescopes [40]. This initial analysis is daunting because of its real-time nature and huge data volumes but can often be effectively performed in parallel on thousands of cores. YouTube videos are primarily meant to be viewed, along with some automated analysis for advertisements or copyright infringements. Twitter data are the subject of intense research in the social sciences [41], especially for topic and sentiment mining, which is performed chiefly on textual "tweets" in the context of associated metadata (e.g., user demographics and temporal information).

Analysis of genomic data involves a more diverse range of approaches because of the variety of steps involved in reading a genome sequence and deriving useful information from it. For population and medical genomics, identifying the genomic variants in each individual genome is currently one of the most computationally complex phases. Variant calling on 2 billion genomes per year, with 100,000 CPUs in parallel, would require methods that process 2 genomes per CPU-hour, three-to-four orders of magnitude faster than current capabilities [42]. Whole genome alignment is another important form of genomic data analysis, used for a variety of goals, from phylogeny reconstruction to genome annotation via comparative methodologies. Just a single whole genome alignment between human and mouse consumes ~100 CPU hours [43]. Aligning all pairs of the ~2.5 million species expected to be available by 2025 amounts to 50–100 trillion such whole genome alignments, which would need to be six orders of magnitude faster than possible today.

Improvements to CPU capabilities, as anticipated by Moore's Law, should help close the gap, but trends in computing power are often geared towards floating point operations and do not necessarily provide improvements in genome analysis, in which string operations and memory management often pose the most significant challenges. Moreover, the bigger bottleneck of Big Data analysis in the future may not be in CPU capabilities but in the input/output (I/O) hardware that shuttles data between storage and processors [44], a problem requiring research into new parallel I/O hardware and algorithms that can effectively utilize them.

The Long Road Ahead

Genomics clearly poses some of the most severe computational challenges facing us in the next decade. Genomics is a "four-headed beast"; considering the computational demands across the lifecycle of a dataset—acquisition, storage, distribution, and analysis—genomics is either on par with or the most demanding of the Big Data domains. New integrative approaches need to be developed that take into account the challenges in all four aspects: it is unlikely that a single advance or technology will solve the genomics data problem. Several key technologies that are most critically needed to support future solutions are discussed in Box 1.

In human health, the major needs are driven by the realization that for precision medicine and similar efforts to be most effective, genomes and related 'omics data need to be shared and compared in huge numbers. If we do not commit as a scientific community to sharing now, we run the risk of establishing thousands of isolated, private data collections, each too underpowered to allow subtle signals to be extracted. More than anything else, connecting these resources requires trust among institutions, scientists, and the public to ensure the collections will be used for medical purposes and not to discriminate or penalize individuals because of their genetic makeup.

Finally, the exascale data and computing centers that are emerging today to meet Big Data challenges in several domains (YouTube [50], Google [51], Facebook [51], and the National Security Agency [52]), are the result of far-sighted planning and commitment by the respective organizations. Now is the time for concerted, community-wide planning for the "genomical" challenges of the next decade.

Box 1. Key Technological Needs for Big Data Genomics

(1) Acquisition

The most important need to sustain the explosive growth in genomic data acquisition is continued advances in sequencing technologies to reduce costs, improve throughput, and achieve very high accuracy. The current costs of ~US$1,000 per human genome begin to make it practical to sequence human genomes in large numbers, especially for critical medical treatments, but to scale to populations of hundreds of millions to billions of genomes, costs must be reduced by at least another one to two orders of magnitude or more. For many medical applications, the time for sequencing must also be reduced so that it can be completed in near real time, especially to rapidly diagnosis acute infections and conditions. Finally, to make a genome sequence most useful, it must be paired with automated methods to collect metadata and phenotype data, all according to appropriate standards so that data collected in one environment can be compared to those collected in another.

(2) Storage

The community needs to start designing and constructing data centers with fast, tiered storage systems to query and aggregate over large collections of genomes and 'omics data. There are new technologies on the horizon that will help support these needs, including 3-D memory, integrated computing technologies that overcome the I/O bottleneck, and networks that are two-to-five orders of magnitude faster because of optical switching [45,46]. Similarly, efficient compression and indexing systems are critical to make the best use out of each available byte while making the data highly accessible. We

also expect algorithmic developments that can represent large collections of personal genomes as a compact graph, making it more efficient and robust to compare one genome to many others. Beyond these approaches, we see the rise of streaming approaches to make on-the-fly comparisons that will allow us to rapidly discard data, especially for sequencing applications that use the sequence data as a means to infer abundances or other molecular activity.

(3) Distribution

The most practical, and perhaps only, solution for distributing genome sequences at a population scale is to use cloud-computing systems that minimize data movement and maximize code federation [47]. New developments from companies such as Google, Amazon, and Facebook that include applications built to fit the frameworks of distributed computing efficient data centers and distributed storage and cloud computing paradigms will be part of the solution. Already, large cloud-based genomic resources are being developed using these technologies, especially to support the needs of the largest sequencing centers or to support the needs of large communities (BGI-cloud, TCGA, the International Cancer Genome Consortium [ICGC], etc.). To make these online systems most useful, the community needs to develop application programming interfaces (APIs) for discovering and querying large datasets on remote systems. The Global Alliance for Genomics and Health [48] and others are beginning to develop such standards for human genomic data, and we expect other communities to follow. Finally, authentication, encryption, and other security safeguards must be developed to ensure that genomic data remain private.

(4) Analysis

Our ultimate goal is to be able to interpret genomic sequences and answer how DNA mutations, expression changes, or other molecular measurements relate to disease, development, behavior, or evolution. Accomplishing this goal will clearly require integration of biological domain expertise, large-scale machine learning systems, and a computing infrastructure that can support flexible and dynamic queries to search for patterns over very large collections in very high dimensions. A number of "data science technologies," including R, Mahout, and other machine learning systems powered by Hadoop and other highly scalable systems, are a start, but the current offerings are still difficult and expensive to use. The community would also benefit from libraries of highly optimized algorithms within a simple interface that can be combined and reused in many contexts as the problems emerge. Data science companies as well as open-source initiatives are already starting to develop such components, such as Amazon's recent "Amazon Machine Learning" prediction system. But because genomics poses unique challenges in terms of data acquisition, distribution, storage, and especially analysis, waiting for innovations from outside our field is unlikely to be sufficient. We must face these challenges ourselves, starting with integrating data science into graduate, undergraduate, and high-school curricula to train the next generations of quantitative biologists, bioinformaticians, and computer scientists and engineers [49].

Supporting Information

S1 Fig. Growth of GenBank. The *y*-axis shows the total sequence in bp. (Blue = GenBank, red = whole genome shotgun [WGS] sequences.) Each line is double of the previous. The *x*-axis indicates time. Each line is 6 months after the previous. Source: http://www.ncbi.nlm.nih.gov/genbank/statistics.
(TIF)

S1 Note. YouTube data estimates.
(DOCX)

S2 Note. Twitter data estimates.
(DOCX)

S3 Note. Human genomic data storage estimates for 2025.
(DOCX)

S4 Note. YouTube distribution statistics (current).
(DOCX)

S1 Table. Capacities of 20 major genomics institutions. The number of sequencers as listed from OmicsMaps.com and their storage capacities from the listed citation. These 20 institutions alone collectively have more than 100 PB of storage available.
(DOCX)

Acknowledgments

We thank V. Jongeneel, M. Baker, and the anonymous reviewers for comments that improved the manuscript and the Illinois CompGen Initiative for providing the collaborative structure that fostered the idea for this analysis. We would also like to thank all of the participants of the 2014 Keystone Symposium on "Big Data in Biology" (organized by Lincoln D. Stein, Doreen Ware, and MCS) for their inspiration for this analysis.

References

1. Mole, B. The gene sequencing future is here. 2014; http://www.sciencenews.org/article/gene-sequencing-future-here.
2. Robinson G.E., et al., Creating a Buzz About Insect Genomes. Science, 2011. 18: 1386.
3. Abecasis G.R., et al., An integrated map of genetic variation from 1,092 human genomes. Nature, 2012. 491(7422): 56–65. doi: 10.1038/nature11632 PMID: 23128226
4. Chin L., Andersen J.N., and Futreal P.A., Cancer genomics: from discovery science to personalized medicine. Nature medicine, 2011. 17(3): 297–303. doi: 10.1038/nm.2323 PMID: 21383744
5. Exome Aggregation Consortium. Exome Aggregation Consortium ExAC Browser. 2015; http://exac.broadinstitute.org/.
6. Giles J., Computational social science: Making the links. Nature, 2012. 488(7412): 448–50. doi: 10.1038/488448a PMID: 22914149
7. Pak, A. and P. Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining, in LREC. 2010. p. 1320–1326.
8. O'Luanaigh, C. Animation shows LHC data processing. 2013; http://home.web.cern.ch/about/updates/2013/04/animation-shows-lhc-data-processing.
9. Newman, R. and J. Tseng. Cloud Computing and the Square Kilometre Array. 2011; http://www.skatelescope.org/uploaded/8762_134_Memo_Newman.pdf.
10. IBM Research. Square Kilometer Array: Ultimate Big Data Challenge. 2013; http://www.skatelescope.org/uploaded/8762_134_Memo_Newman.pdf.

11. Omicsmap. Next Generation Genomics: World Map of High-throughput Sequencers. 2015; http://omicsmaps.com/.
12. Regalado, A. EmTech: Illumina Says 228,000 Human Genomes Will Be Sequenced This Year. MIT Technology Review 2014 [cited 2015 April 28, 2015]; http://www.technologyreview.com/news/531091/emtech-illumina-says-228000-human-genomes-will-be-sequenced-this-year/.
13. AllSeq. HiSeq X Ten. 2015 [cited 2015 April 1, 2015]; http://allseq.com/x-ten.
14. Illumina. HiSeq X Series of Sequencing Systems. 2015 [cited April 28, 2015]; http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-x-ten.pdf.
15. Mora C., et al., How Many Species Are There on Earth and in the Ocean? PLoS Biology, 2011; 9: e1001127. doi: 10.1371/journal.pbio.1001127 PMID: 21886479
16. Li J.Y., Wang J., and Zeigler R.S., The 3,000 rice genomes project: new opportunities and challenges for future rice research. GigaScience, 2014. 3: 8. doi: 10.1186/2047-217X-3-8 PMID: 24872878
17. Zhu J., A year of great leaps in genome research. Genome medicine, 2012. 4(1): 4. doi: 10.1186/gm303 PMID: 22293069
18. Eisen J.A., Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS Biology, 2007. 5(3): e82. PMID: 17355177
19. Gevers D., et al., The Human Microbiome Project: a community resource for the healthy human microbiome. PLoS Biology, 2012. 10(8): e1001377. doi: 10.1371/journal.pbio.1001377 PMID: 22904687
20. Genomics England. The 100,000 Genomes Project. 2015; http://www.genomicsengland.co.uk/the-100000-genomes-project/.
21. Briggs, H (2013) Hundred thousand genomes to be mapped in Saudi Arabia. BBC News. http://www.bbc.com/news/health-25216135
22. Sulem P., et al., Identification of a large set of rare complete human knockouts. Nature genetics, 2015; 47: 448–452. doi: 10.1038/ng.3243 PMID: 25807282
23. Kaiser, J. White House fleshes out Obama's $215 million plan for precision medicine. Science Insider 2015; http://news.sciencemag.org/biology/2015/01/white-house-fleshes-out-obama-s-215-million-plan-precision-medicine.
24. Internet World Stats. 2015; http://www.internetworldstats.com/stats.htm.
25. Soon W.W., Hariharan M., and Snyder M.P., High-throughput sequencing for biology and medicine. Molecular systems biology, 2013. 9: 640. doi: 10.1038/msb.2012.61 PMID: 23340846
26. Chen R., et al., Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell, 2012. 148(6): 1293–307. doi: 10.1016/j.cell.2012.02.009 PMID: 22424236
27. Hsi-Yang Fritz M., et al., Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome research, 2011. 21(5): 734–40. doi: 10.1101/gr.114819.110 PMID: 21245279
28. Loh P.-R., Baym M., and Berger B., Compressive genomics. Nature Biotechnology, 2012. 30: 627–630. doi: 10.1038/nbt.2241 PMID: 22781691
29. Koren S. and Phillippy A.M., One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Current opinion in microbiology, 2015. 23C: 110–120.
30. Christley S., et al., Human genomes as email attachments. Bioinformatics, 2009. 25(2): 274–5. doi: 10.1093/bioinformatics/btn582 PMID: 18996942
31. Patro R., Mount S.M., and Kingsford C., Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature biotechnology, 2014. 32(5): 462–4. doi: 10.1038/nbt.2862 PMID: 24752080
32. Golden A., Djorgovski S., and Greally J., Astrogenomics: big data, old problems, old solutions? Genome Biology, 2013. 8: 129.
33. Djoerd Hiemstra C.H., Brute Force Information Retrieval Experiments using MapReduce'. ERCIM News, 2012. 89: 31–32.
34. Smith, B., Data Transport for the Square Kilometre Array, in UbuntuNet Alliance Annual Conference. 2012. 15–22.
35. Global Internet Phenomena Report. 2013; http://www.sandvine.com/downloads/general/global-internet-phenomena/2013/2h-2013-global-internet-phenomena-report.pdf.
36. Sboner A., et al., The real cost of sequencing: higher than you think! Genome Biology, 2011. 12: 125. doi: 10.1186/gb-2011-12-8-125 PMID: 21867570
37. Marx, V., Drilling into big cancer-genome data. Nature Methods, 2013. 10: p. 293–297.
38. Baker M., Next-generation sequencing: adjusting to data overload. Nature Methods, 2010. 7: 495–499.

39. Erlich Y. and Narayanan A., Routes for breaching and protecting genetic privacy. Nature reviews. Genetics, 2014. 15(6): 409–21. doi: 10.1038/nrg3723 PMID: 24805122

40. Norris, R.P., Data Challenges for Next-generation Radio Telescopes. 2011.

41. Kumar S., Morstatter F., and Liu H., Twitter Data Analytics. 2014: Springer.

42. Langmead B., et al., Searching for SNPs with cloud computing. Genome Biol, 2009. 10(11): R134. doi: 10.1186/gb-2009-10-11-r134 PMID: 19930550

43. Kurtz S., et al., Versatile and open software for comparing large genomes. Genome Biol, 2004. 5(2): R12. PMID: 14759262

44. Trelles O., et al., Big data, but are we ready? Nature reviews. Genetics, 2011. 12(3): 224.

45. Loh, G.H. 3D-stacked memory architectures for multi-core processors. in ACM SIGARCH. 2008.

46. Chan, V.W., et al. Optical flow switching: An end-to-end "UltraFlow" architecture. in 15th International Conference on Transparent Optical Networks (ICTON). 2013. IEEE. doi: 10.1109/ICTON.2013.6602704

47. Schatz M.C., Langmead B., and Salzberg S.L., Cloud computing and the DNA data race. Nature biotechnology, 2010. 28(7): 691–3. doi: 10.1038/nbt0710-691 PMID: 20622843

48. Global Alliance for Genomics and Health. Global Alliance for Genomics and Health. 2015 April 28, 2015]; http://genomicsandhealth.org/.

49. Schatz, M.C., Computational thinking in the era of big data biology. Genome Biology, 2012. 13(11): 177.

50. Hollis, C. EMC's Record Breaking Product Launch. 2011 April 28, 2015]; http://chucksblog.emc.com/chucks_blog/2011/01/emcs-record-breaking-product-launch.html.

51. Hoff, T. How Google Backs up the Internet along with Exabytes of other data. 2014; http://highscalability.com/blog/2014/2/3/how-google-backs-up-the-internet-along-with-exabytes-of-othe.html.

52. Daily Kos. Utah Data Center stores data between 1 exabyte and 1 yottabyte. 2013; http://www.dailykos.com/story/2013/08/05/1228923/-Utah-Data-Center-stores-data-between-1-exabyte-and-1-yottabyte.

Concepts of Precision Medicine in Breast Cancer

An Expert Interview with Eleni Andreopoulou

Weill Cornell Medicine, Division of Hematology and Medical Oncology, Weill Cornell Breast Center, New York, NY, US

DOI: https://doi.org/10.17925/OHR.2018.14.1.16

Eleni Andreopoulou

Dr Eleni Andreopoulou is an Associate Professor of Medicine and the Director of Breast Cancer Clinical Research at Weill Cornell Medicine and New York-Presbyterian Hospital, New York, NY, US, where she specializes in the care and treatment of patients with breast cancer. She is also a member of the Englander Institute for Precision Medicine at Cornell. She previously held faculty positions at the Albert Einstein College of Medicine and Montefiore Medical Center, New York, and the University of Texas MD Anderson Cancer Center, Houston, Texas, US. Dr Andreopoulou completed her training in major academic institutions in both Europe and the US, including St Bartholomew's Hospital and the Royal Marsden Hospital and Institute of Cancer Research in London, UK, and the New York University School of Medicine. She was a European Society of Medical Oncology fellow and was also awarded the Calabresi Scholarship in mentored cancer therapeutics. Dr Andreopoulou has a special interest in the individualization of patient treatment, particularly in caring for women with aggressive breast cancer. Her main research interest involves precision medicine to fast-track the drug development of biologics and targeted therapy to effectively manage, treat and cure breast cancer. She is involved with all phases of clinical drug development and especially focuses on innovative preoperative clinical trial design that incorporates cutting-edge technology. Dr Andreopoulou's research involves projects focused on pharmacogenomics predictors of response to treatment for early and advanced stage breast cancer. Dr Andreopoulou is an active investigator of several clinical trials of novel therapeutic approaches in advanced disease, including her leadership role as a principal investigator in the development of drugs sponsored by the Cancer Therapy Evaluation Program at the National Cancer Institute. Dr Andreopoulou actively facilitates the interface between basic and applied research at the Meyer Cancer Center at Weill Cornell Medicine. She leads a multidisciplinary, prospective breast cancer biobank to provide a crucial foundation for precision medicine research. Dr Andreopoulou has also been active with breast cancer awareness programs covering screening and prevention with a particular focus in serving underserved minorities in the local area. She has published several peer-reviewed articles, reviews, editorials and book chapters. She is a member of the American Society of Clinical Oncology, the American Association of Cancer Research, the American Women`s Medical Association the Royal Society of Medicine in England, and the European Society of Medical Oncology.

Keywords

Breast cancer, precision medicine, neoadjuvant treatment, inta-tumor heterogeneity, snapshots of clonal evolution, preclinical models, fast-track drug development

Disclosure: Eleni Andreopoulou has nothing to disclose in relation to this article.

Review Process: This is an expert interview and, as such, has not undergone the journal's standard peer review process.

Authorship: All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship of this manuscript, take responsibility for the integrity of the work as a whole, and have given final approval to the version to be published.

Received: April 18, 2018

Published Online: April 27, 2018

Citation: *Oncology & Hematology Review*. 2018;14(1):16–17

Corresponding Author: Eleni Andreopoulou, Weill Cornell Medicine, Division of Hematology and Medical Oncology, Weill Cornell Breast Center, 425 East 61st Street, 8th Floor, New York, NY 10065, US. E: ela9082@med.cornell.edu
Twitter: @eleniandreop
Facebook: @CornellBreastCenter

Support: No funding was received in the publication of this article.

Q. Could you tell us a little about the latest research in organoids and biomimetic platforms in precision medicine?

Precision medicine aims to utilize information about a patient's tumor, including gene alterations that may aid in the identification of effective therapies.[1] Recent advances have allowed researchers to combine genomic analysis with *ex vivo* drug screens. The opportunity to develop preclinical models that retain the tumor's basic characteristics provides a platform for biomarker discovery and high-throughput drug screening. Patient-derived tumor xenografts have emerged as powerful systems to study many cancer types by growing tumor cells in immunocompromised mice. Advances in this area have focused on trying to improve engraftment rates and introduce the patient's own immune cells.[2] For more rapid and less costly screens, the patient's tumor cells are now also cultured in the laboratory in 3D as organoids, allowing us to screen thousands of candidate drugs and drug combinations that have the potential to be used in the clinic.[3] The future of personalized medicine resides in being able to incorporate cells of the microenvironment in state-of-the-art biomimetic platforms. This method of screening will better account for the influence of the cells that are adjacent to the tumor, known to influence the growth of cancers and their response to treatment.[4]

Q. What role may artificial intelligence play in precision medicine?

Analysis of large data sets is a major driver in the development of precision medicine. Transforming big data into clinically-actionable knowledge can be a tedious process, requiring specialized personnel and efficient computational methods. Artificial intelligence (AI) and more specifically machine learning (ML) can play a key role in providing clinicians with faster access to medical data (*Precision Medicine Knowledge Base AI Bot* – WCM/NYP collaboration with Microsoft),[5] better diagnostic tools,[6,7] and drug discovery methods.[8] Furthermore, integrating precision medicine data into electronic health records (EHRs)[9] and applying AI and ML methods to EHRs will allow clinicians to better match patients to novel targeted therapies by exploiting the molecular vulnerabilities of their disease.

Q. How can we address the challenges related to security and privacy with personal health information?

In an era characterized by data breaches and the inappropriate sharing of personal information online, it is increasingly important to ensure that patient health and privacy information is adequately protected. Adhering to Health Insurance Portability and Accountability Act (HIPAA) guidelines and adopting the latest security standards is a great first step to ensure that both security and privacy with personal health information are protected. Building on top of HIPAA, the National Institute of Health's All of Us Research Program[10] (formerly known as Precision Medicine Initiative) has established the *PMI Privacy and Trust Principles*[11] and the *PMI Data Security Policy Principles and Framework*.[12] These principles promote transparency, respect of participant preferences, and describe in detail how data are shared, accessed, used, and how data quality and integrity is maintained. Furthermore, they outline how to protect, detect, respond to, and recover from a malicious attack from a third-party actor.

Q. What have neoadjuvant endocrine therapy studies taught us about the role of precision medicine in breast cancer?

Neoadjuvant endocrine therapy provides a unique opportunity to study endocrine-sensitive and -resistant breast cancer with hormone receptor-positive phenotype. This setting is gaining traction for accelerated development of effective therapy allowing integration of biomarkers and surrogate endpoints into the process of care for tumors that exhibit endocrine therapy resistance. Molecular testing at diagnosis to define the genetic "fingerprint" and accompanying molecular dependencies of the tumors we seek to eliminate, followed by longitudinal assessments of both clinical and biomarker responses, allows for patient selection to enroll in novel clinical trials exploring the impact of agents that aim to enhance response beyond that of endocrine treatments alone. From the precision medicine standpoint several lessons have been learned: the slow emergence of downstaging is relating to lower rate of apoptosis with endocrine therapy and dependence of response on the antiproliferative effects of estrogen deprivation. While change in Ki67 is accepted as a validated endpoint for comparing endocrine neoadjuvant agents, on-treatment levels of Ki67 are related to long-term prognosis more closely than pretreatment Ki67. The Preoperative Endocrine Prognostic Index (PEPI) combines residual Ki67 score with measures of on-treatment estrogen receptor (ER) and other clinicopathologic factors and has clinical utility. Preliminary studies demonstrate that tumors that exhibit aromatase inhibitor resistant proliferation in the neoadjuvant setting is often sensitive to cyclin-dependent kinase inhibitor (CDKi) CDK4/6i. Serial Ki67 monitoring before surgery is therefore the logical approach to tailored use of adjuvant CDK4/6i treatment.[13] Recent data demonstrated more pronounced treatment effect with the phosphoinositide 3-kinase (PI3K) selective inhibitor in the PIK3CA-mutated hormone receptor-positive, human epidermal growth factor receptor 2 (HER2)-negative breast cancer.[14] Supression of proliferation is a pharmacodynamic indicator of response, but possibly not the only one in the ER signaling pathway. A more comprehensive study of residual disease will elucidate the constitute of response. Although many challenges remain to be addressed, the implementation of molecular testing has introduced new possibilities for informing precision medicine decisions. The Trial of Perioperative Endocrine Therapy – Individualizing Care (POETIC) is ongoing (NCT02338310).[15]

Q. How can precision medicine address the challenge of intra-tumor heterogeneity resulting from clonal evolution of the disease?

Advances in genetic sequencing analysis are helping to illuminate the prevalence of intra-tumor heterogeneity, which allows us to answer compelling high priority clinical questions. Intra-tumor heterogeneity refers to cellular diversity attributed to genetic and epigenetic factors, and to non-hereditary adaptive responses to selective pressures through the dynamic evolution of cancerous clones. Intra-tumor heterogeneity poses significant challenges and rationalizes differences in response to treatment.[16]

Precision medicine aims to decode the intra-tumor heterogeneity.[16] The opportunity to define and re-define the molecular signature of a given tumor at multiple time points along its evolutionary lineage enables an understanding of the clonal evolution of tumors. Beyond untreated tumors, the ability to perform whole exome sequencing and clonality analysis of metastatic tumors is crucial to detect potentially actionable mutagenic divergence due to dynamic clonal evolution through the natural course of the tumor and treatment effect. The recent development of single-cell sequencing tools allows the transcriptomes of thousands of cells to be processed simultaneously in order to identify subpopulations of cells and provide functional insights.[17,18] Understanding how tumors that are inherently chemoresistant and hard wired to migrate and invade as well as treatment influence in directing the evolution of different subclones is a compelling question with significant clinical implications, and its answer will allow therapy to be tailored to a changing tumor and its microenvironment. In this setting we clinicians work very closely with our scientist colleagues at the Englander Institute of Precision Medicine at Weill Cornell (https://eipm.weill.cornell.edu). ❑

1. Beltran H, Eng K, Mosquera JM, et al. Whole-exome sequencing of metastatic cancer and biomarkers of treatment response. *JAMA Oncol*. 2015;1:466–74.
2. Kyriakides PW, Inghirami G. Are we ready to take full advantage of patient-derived tumor xenograft models? *Hematol Oncol*. 2018;36:24–7.
3. Pauli C, Hopkins BD, Prandi D, et al. Personalized *in vitro* and *in vivo* cancer models to guide precision medicine. *Cancer Discov*. 2017;7:462–77.
4. Toyoda Y, Celie KB, Xu JT, et al. A 3-dimensional biomimetic platform to interrogate the safety of autologous fat transfer in the setting of breast cancer. *Ann Plast Surg*. 2018. Doi: 10.1097/SAP.0000000000001364.
5. Nixon J, DevRadio, Caldwell C, et al. Behind the scenes: How Weill Cornell Medicine built a chatbot for clinicians to gain fast access to medical data. 2017. Available at: https://channel9.msdn.com/Blogs/DevRadio/DR1747 (accessed April 18, 2018).
6. Khosravi P, Kazemi E, Imielinski M, et al. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EbioMedicine*. 2018;27:317–28.
7. Motlagh NH, Jannesary M, Aboulkheyr HR, et al. Breast cancer histopathological image classification: A deep learning approach. *BioRxiv*. 2018. DOI: https://doi.org/10.1101/242818.
8. Madhukar NS, Elemento O. Bioinformatics approaches to predict drug responses from genomic sequencing. *Methods Mol Biol*. 2018;1711:277–96.
9. GenomeWeb Clinical Sequencing. Weill Cornell Integrates Cancer Exome Test into EHRs With Eye Toward Next Version, 2018. Available at: www.genomeweb.com/informatics/weill-cornell-integrates-cancer-exome-test-ehrs-eye-toward-next-version#.WtczdYhuY3s (accessed April 18, 2018).
10. National Institutes of Health. All of Us Research Program, 2018. Available at: https://allofus.nih.gov/ (accessed April 18, 2018).
11. The White House. Precision Medicine Initiative: Privacy and Trust Principles. 2015. Available at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/finalpmiprivacyandtrustprinciples.pdf (accessed April 18, 2018).
12. The White House. Precision Medicine Initiative: Data Security Policy Principles and Framework. 2016. Available at: https://obamawhitehouse.archives.gov/sites/obamawhitehouse.archives.gov/files/documents/PMI_Security_Principles_Framework_v2.pdf (accessed April 18, 2018).
13. Ma CX, Gao F, Luo J, et al. NeoPalAna: Neoadjuvant Palbociclib, a cyclin-dependent kinase 4/6 inhibitor, and anastrozole for clinical stage 2 or 3 estrogen receptor-positive breast cancer. *Clin Cancer Res*. 2017;23:4055–65.
14. ESMO. ESMO 2017 Press Release: LORELEI: Taselisib Boosts Breast Tumor Shrinkage. 2017. Available at: www.esmo.org/Conferences/Past-Conferences/ESMO-2017-Congress/Press-Media/Press-Releases/LORELEI-Taselisib-Boosts-Breast-Tumor-Shrinkage (accessed April 18, 2018).
15. Clinicaltrials.gov. Trial of Perioperative Endocrine Therapy - Individualising Care (POETIC). ClinicalTrials.gov Identifier: NCT02338310. Available at: https://clinicaltrials.gov/ct2/show/NCT02338310 (accessed April 24, 2018).
16. Martelotto LG, Ng CKY, Puscuoglio S, et al. Breast cancer intra-tumor heterogeneity. *Breast Cancer Res*. 2014;16:210.
17. Redmond D, Poran A, Elemento O. Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. *Genome Med*. 2016;8:80.
18. Kidess E, Jeffrey SS. Circulating tumor cells versus tumor-derived cell-free DNA: rivals or partners in cancer care in the era of single-cell analysis? *Genome Med*. 2013;5:70.