bioRxiv preprint doi: https://doi.org/10.1101/554394; this version posted February 19, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RESEARCH

Multi-omic analysis supports a developmental hierarchy of molecular subtypes in high-grade serous ovarian carcinoma

Ludwig Geistlinger[1,2], Seyhun Oh[1,2], Marcel Ramos[1,2,3], Lucas Schiffer[1,2], Boris Winterhoff[4,5], Martin Morgan[3], Giovanni Parmigiani[6], Michael Birrer[7], Li-Xuan Qin[8], Markus Riester[9], Tim Starr[4,5] and Levi Waldron[1,2*]

Abstract

Background: The majority of ovarian carcinomas are of high-grade serous histology, which is associated with poor prognosis and limited treatment options. Several studies have identified gene expression-based subtypes of high-grade serous ovarian carcinoma (HGSOC) as a basis for targeted therapy, yet extensive ambiguity in subtype classification impairs translation of these subtypes into clinical practice. Furthermore, although HGSOC tumors are known to be frequently polyclonal, it is unknown whether clones within the same tumor share the same subtype.

Results: We investigate whether ambiguity in subtype classification can be attributed to the polyclonal composition of HGSOC tumors, addressing the currently unresolved question whether proposed subtypes are early or late events in tumorigenesis. This hypothesis is first tested in The Cancer Genome Atlas HGSOC cases by (i) analyzing recurrent somatic copy number alterations for their association with subtypes, (ii) inferring per-alteration clonality from complementary analysis of SNP arrays and whole-exome sequencing, and (iii) testing whether subtype-associated alterations tend to predominantly occur clonally (early events) or subclonally (late events). As opposed to the genomically distinct evolution of soft-tissue sarcoma subtypes, we find that subtype association of HGSOC alterations significantly correlate with subclonality. This correlation is particularly evident for the high-purity proliferative subtype spectrum, which is also characterized by extreme genomic instability, absence of immune infiltration, and increased patient age. This is in stark contrast to the high-purity differentiated subtype spectrum, which is characterized by largely intact genome integrity, high immune infiltration, and younger patient age. Other subtypes showed intermediate levels for these characteristics. From single cell sequencing of an independent HGSOC tumor, we demonstrate that ambiguity in subtype classification extends to individual tumor epithelial cells, further supporting a developmental transition from one subtype spectrum to another.

Conclusion: We propose a novel model of HGSOC tumor development that complements the subtype perspective. In this model, individual tumors develop from an early differentiated spectrum to a late proliferative spectrum, and may exhibit characteristics of different previously defined "subtypes" at different points along a timeline characterized by increasing genomic instability and subclonal expansion. This model is more consistent with available bulk and single-cell data, and provides an explanation for ambiguity in subtype classification as the result of assigning discrete, mutually exclusive subtypes to a genomically complex process of tumor evolution.

Keywords: ovarian cancer; tumor evolution; intra-tumor heterogeneity; absolute copy number; transcriptome subtypes; single-cell sequencing

Introduction

High-grade serous ovarian cancer (HGSOC) is a genomically complex disease, for which the accurate characterization of molecular subtypes is difficult but

*Correspondence: Levi.Waldron@sph.cuny.edu

[1]Graduate School of Public Health and Health Policy, City University of New York, 55 W 125th St, New York, NY 10027, USA

Full list of author information is available at the end of the article

bioRxiv preprint doi: https://doi.org/10.1101/554394; this version posted February 19, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

anticipated to improve treatment and clinical outcome. Several previous studies have devoted substantial research effort to identifying molecularly distinct HGSOC subtypes by clustering tumors together that have similar overall transcriptome profiles [1–5]. Among these studies, The Cancer Genome Atlas (TCGA) project reported four subtypes, and termed these *differentiated*, *immunoreactive*, *mesenchymal*, and *proliferative* [2]. These names are based on marker gene expression and have been adopted by subsequent subtyping efforts.

However, robustness and clinical utility of these subtypes remain controversial [6, 7]. Based on a compendium of 15 microarray datasets consisting of ≈1,800 HGS ovarian tumors, corresponding subtype classifiers identified subsets significantly differing in overall survival, but were not robust to re-fitting in independent datasets and grouped only approximately one third of patients concordantly into four subtypes [8].

This ambiguity in tumor classification might arise from an intra-tumor admixture of different subtypes. Recent studies have indicated that most HGS ovarian tumors are polyclonal, meaning that a single tumor is a heterogeneous assembly of distinct cancer genotypes arising from different subclones. Lohr *et al.* estimated that 95% of the tumors in the TCGA HGSOC dataset are polyclonal, and ≈40% consist of ≥4 subclones [9]. Given the extensive polyclonality of HGS tumors, we hypothesize that subtype assignment via transcriptome clustering is biased towards late events. Identified subtypes will thus be subclone-specific, making transcriptome clustering unlikely to provide patient subsets that will benefit from specific treatment. If a tumor contains multiple subclones that are classified as different subtypes, subtype-specific treatments will only be effective against selected subclones within a single tumor.

To test this hypothesis, we focus on somatic copy number alterations (SCNAs) given their known causal roles in oncogenesis and their reported potential to discriminate between cancer types and subtypes [10–12]. The `GISTIC2` method detects SCNAs that are more recurrent than expected by chance, in order to distinguish cancer-driving events from random passenger alterations [13]. The method also separates *arm-level* events, defined as broad SCNAs covering a large fraction of a chromosome arm, from *focal* events of relatively small range. Recurrent focal SCNAs have repeatedly been shown to harbor known oncogenes and tumor suppressor genes [14, 15].

The `ABSOLUTE` algorithm infers tumor purity and ploidy directly from the analysis of SCNAs [16]. Accounting for the intermixture of cancer cells with normal cells within a tumor sample (*purity*), and the often abnormal DNA content of cancer cells (*ploidy*), is crucial for the accurate quantification of an alteration's absolute copy number per cancer cell. Furthermore, it also allows to identify SCNAs not fitting a tumor's purity and ploidy relationship as a consequence of subclonal evolution.

Leveraging publicly available `GISTIC2` and `ABSOLUTE` SCNA calls in TCGA HGSOC tumors, we analyze whether recurrent subtype-associated copy number alterations display greater intra-tumor heterogeneity than other alterations. We assess the reliability of these calls by absolute copy number analysis on whole-exome sequencing data, and complement results with single-cell subtype classification on an independent HGSOC tumor.

Results

We previously reported a systematic assessment of the four reported HGSOC subtypes (*differentiated*, *immunoreactive*, *mesenchymal*, and *proliferative*) with respect to robustness and association to overall survival [8]. Based on 1,774 HGSOC tumors from 12 studies available in the `curatedOvarianData` database [17], we found that only a minority of the tumors can be classified robustly.

Here, we test the hypothesis that the observed ambiguity in tumor classification is a consequence of intra-tumor heterogeneity (Figure 1). For the purpose of testing this hypothesis, we focus on the subtypes proposed by TCGA [2], and integrate information as available for 516 TCGA HGS ovarian tumors.

Subtype purity, ploidy, and subclonality

Previous studies reported specific clinical and tumor pathology characteristics of the four subtypes [2, 8]. Using per-tumor estimates as obtained with `ABSOLUTE`, we observed significant differences in tumor purity between subtypes (Supplementary Figure **??**a). In particular, tumors of differentiated subtype are characterized by high purity, but significantly lower ploidy and subclonality than the other three subtypes (Supplementary Figure **??**b,c). Lower ploidy for tumors of differentiated subtype was in agreement with a significantly lower number of genome doublings (Supplementary Figure **??**d).

Subtype association of recurrent SCNAs

We next analyzed recurrent focal SCNAs as identified with `GISTIC2` in TCGA HGS ovarian tumors for association with the four subtypes (Figure 2). We tested a total of 70 recurrent focal SCNAs comprising 31 amplifications and 39 deletions (Figure 2, outer ring). Nominal p-values for the 70 focal alterations showed a concentration of p-values near zero (Supplementary

bioRxiv preprint doi: https://doi.org/10.1101/554394; this version posted February 19, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure **??**a), corresponding to 35 of 70 alterations being significantly associated with subtypes (FDR < 0.1, Figure 2, inner ring).
Associations with the proliferative subtype were significantly overrepresented among the subtype-associated regions (20 out of 35, $p = 0.007$, Fisher's exact test, Figure 2, barplot). Regions of strongest subtype association included the *FRS2*-containing amplification on chromosome 12 and the *BLC2L1*-containing amplification on chromosome 20 (Figure 2, gene names).

Correlation of subtype association with subclonality

We test the hypothesis that the reported HGSOC subtypes differentiate late in tumorigenesis by assessing the correlation between subtype association and subclonality of recurrent CN alterations (Figure 1A-E). Subtype association of an alteration is calculated via a score S_A, corresponding to the χ^2 test statistic (Figure 3A). Subclonality of an alteration is calculated via a score S_C, defined as the fraction of samples for which this alteration is subclonal (Figure 3B). See also Methods for details on how both scores are calculated.
Under the null hypothesis that subtype-associated alterations occur no earlier or later than other alterations, Spearman correlation ρ between S_A and S_C would be expected to be zero:

$$H_0: \quad \rho(S_A, S_C) = 0 \tag{1}$$

Rejection of H_0 has clear interpretation: if subtype-associated SCNAs tend to be subclonal, i.e.

$$\rho(S_A, S_C) > 0, \tag{2}$$

this suggests that the subtypes are late events in tumor evolution. If subtype-associated alterations tend not to be subclonal, i.e.

$$\rho(S_A, S_C) < 0, \tag{3}$$

this would suggest that subtypes are early events, consistent with these being intrinsic subtypes.
As illustrated in Figure 3C, we obtained a significant positive correlation between subtype association and subclonality of the 70 recurrent focal SCNAs depicted in Figure 2. To account for non-independence of the occurrence of different SCNAs, we also carried out a permutation test, which confirmed the significance of this finding ($p = 0.006$). When stratifying tumors by purity to assess the possibility of confounding, the correlation was positive in all strata and did not significantly differ between strata (Supplementary Figure **??**). Regions of highest subtype association and subclonality comprised throughout amplifications, and repeatedly displayed increased alteration frequency for the proliferative subtype (including the *BRD4* amplification and the telomeric 20q13.33 amplification shown in Figure 4 bottom left; additional regions shown in Supplementary Figure **??**). A notable exception was the highly subconal *MYC*-containing amplification on chromosome 8, which displayed decreased alteration frequency for tumors of proliferative subtype as previously reported [2]. In contrast, predominantly clonal alterations were enriched for deletions (9 of 10 regions with $S_C < 0.3$) with comparatively moderate subtype association (including loss of *PPP2R2A* and *MGA* as shown in Figure 4 top left). In agreement with previous studies that reported frequent loss of *PTEN*, *RB1*, and *NF1* in HGS ovarian tumors [2, 18], we also observed alterations in these regions to occur predominantly clonal and largely irrespective of subtype classification (Supplementary Figure **??**).

Soft tissue sarcoma as a negative control

HGS ovarian carcinoma and adult soft tissue sarcoma (STS) are both characterized by low levels of somatic mutations, but high levels of SCNAs [19]. In contrast to TCGA ovarian carcinomas which are exclusively of high-grade serous type, TCGA ST sarcomas represent several major types each characterized by specific genomic features as expected for true intrinsic subtypes in the sense of Equation 3.
Transcriptome clustering of 259 TCGA ST sarcomas [19] was largely determined by STS type with (i) one cluster exclusively composed of Leiomyosarcoma (LMS), (ii) another cluster dominated by dedifferentiated liposarcoma (DDLPS), and (iii) the third cluster mostly consisting of undifferentiated pleomorphic sarcoma (UPS) and myxofibrosarcoma (MFS), two molecularly closely related STS types [20].
When testing a total of 64 recurrent focal SCNAs (23 amplifications / 41 deletions) for association with the three transcriptome clusters, we found a strong enrichment of nominal p-values near zero, corresponding to 41 of 60 alterations being significantly associated (FDR < 0.1, Supplementary Figure **??**b). Association with the STS type-dominated transcriptome clusters was negatively correlated with subclonality of the 64 SCNAs (Figure 3D), consistent with the assumption of these being intrinsic STS type-specific events. Regions of strongest subtype association and concomitantly low subclonality included the *MDM2* amplification (Figure 4 top right), previously reported to be a key driver of DDLPS and MFS/UPS, but rarely occurring in LMS [19, 21]. A notable exception from the observed trend was the *TP73*-containing telomeric deletion on chromosome 1 that occurred predominantly subclonal in sarcomas assigned to the DDLPS and MFS/UPS clusters, yet predominantly clonal in the LMS cluster (Figure 4 bottom right).

bioRxiv preprint doi: https://doi.org/10.1101/554394; this version posted February 19, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Consistency with whole-exome sequencing

To establish the reliability of inferring SCNA subclonality with `ABSOLUTE` from SNP-array data, we next investigated whether results are consistent when using whole-exome sequencing data instead [22]. We applied `PureCN` [20], which, conceptually similar to `ABSOLUTE`, takes tumor purity and ploidy into account, but is optimized for SCNA calling from targeted short read sequencing data. As reported elsewhere [22], per-tumor estimates of purity and ploidy were in good agreement between platforms (Pearson correlation of 0.77 for purity and 0.74 for ploidy). This also applied to individual copy number calls when analyzed in recurrent `GISTIC2` regions, where we found a median concordance of 87.7% (corresponding to the percentage of tumors with identical CN state for one `GISTIC2` region at a time).

Evidence from single-cell sequencing

We next analyzed whether ambiguity in tumor classification arises from ambiguity on the cellular level, or as a result of confidently classifying individual tumor cells as different subtypes (Figure 1F). We therefore applied the `consensusOV` classifier that was trained on concordantly classified tumors by three major subtype classifiers across 15 microarray datasets [8]. Notably, the consensus classifier also displayed high concordance when comparing classification on RNA-seq data and microarray data for TCGA HGS ovarian tumors (Figure 5A).

Figure 5C shows the resulting subtype calls when applying the consensus classifier to 66 cells of an HGS ovarian tumor for which a recent study reported heterogeneity within ovarian cancer epithelium and cancer associated stromal cells [23]. The majority of epithelial cells (33 of 37) were classified as immunoreactive, in agreement with the classification of the bulk tumor (IMR: 0.646, DIF: 0.164, MES: 0.142, PRO: 0.048). Several cells (8 of 29) assigned to the stromal group by Winterhoff *et al.* [23] were classified as mesenchymal, which we and others [24] found before to be a low-purity subtype (Supplementary Figure **??**a).

Classification margin scores, i.e. the difference between the top two subtype scores, were systematically lower for individual cells (0.239 ± 0.151) than for the bulk tumor (0.482, Figure 5B). Inspecting individual subtype classification probabilities of epithelial cells classified as immunoreactive (*ClassProb* bars for each subtype in Figure 5C) thereby revealed the differentiated subtype to often closely placing second. Vice versa, the four epithelial cells classified as differentiated had the immunoreactive subtype closely placing second.

To analyze whether the observed ambiguity in classification of single cells can be solely explained by zero-inflation of scRNA-seq data, rendering parts of the 100-gene signature of the consensus classifier not sufficiently informative, we also used an extended signature of 800 genes (see Methods). However, highly similar subtype calls (Supplementary Figure **??**) and margin score distribution (Figure 5B) indicated that the 100-gene signature already sufficiently captures subtype-specific expression on the level of single cells. In a complementary analysis, we downsampled the TCGA bulk RNA-seq data to match the coverage of the scRNA-seq data. Classification margin scores on the downsampled data closely resembled the distribution observed on the original data, and clearly exceeded the range of margin scores observed on the scRNA-seq data (Figure 5B).

Discussion

We analyzed HGSOC subtypes in the context of intra-tumor heterogeneity and investigated whether ambiguity in subtype classification can be attributed to the polyclonal composition of HGSOC tumors, addressing the currently unresolved question whether proposed subtypes are early or late events in tumorigenesis. We therefore (i) analyzed recurrent focal SCNAs for association with subtypes, and (ii) tested whether subtype-associated SCNAs tend to predominantly occur clonally (early events) or subclonally (late events).

From subtype association analysis, we found a large fraction of recurrent SCNAs detected in TCGA HGS ovarian tumors to be associated with subtypes. Association with the proliferative subtype was significantly over-represented, which comprised a disproportional large fraction of CN amplifications. This was in line with an overall higher ploidy and increased frequency of genome duplication for tumors of proliferative subtype.

Association of SCNAs with subtypes was positively correlated with subclonality of SCNAs, particularly driven by alterations associated with the proliferative subtype such as amplifications of *BCL2L1*, *BRD4*, and *MYC*. Closer inspection of individual SCNAs repeatedly displayed decreased alteration frequency with relatively small subclonal fractions for tumors of differentiated subtype, as opposed to increased alteration frequency with relatively large subclonal fractions for the proliferative subtype. The diametral behavior of the differentiated and the proliferative subtype, both comprising tumors of high purity, was also evident in a close-to-normal ploidy and a small subclonal genome fraction of the differentiated subtype.

A subtype model based on HGSOC tumor evolution: our observations are consistent with a model that places the differentiated and the proliferative subtype at opposite ends of the timeline of HGSOC tumor developement; with the differentiated subtype being an early subtype, the proliferative a late subtype, and

bioRxiv preprint doi: https://doi.org/10.1101/554394; this version posted February 19, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the immunoreactive and the mesenchymal being intermediate subtypes. HGSOC development along this timeline is thereby reportedly characterized by an increasing level of genomic instability and subclonal expansion [25, 26].

Several previous findings support this model: (i) mean age at diagnosis was lowest for patients of differentiated subtype, but significantly increased for patients of proliferative subtype [8], and (ii) tumors of differentiated subtype displayed a high level of infiltrating immune cells, indicating an active immune response at an early time point of tumor development, whereas tumors of proliferative subtype displayed a negligible level of infiltrating immune cells, consistent with an adapted tumor successfully evading the immune response at a late point in tumor evolution [27]. It seems initially counterintuitive that the proliferative subtype displayed a lower risk than the mesenchymal subtype with respect to overall survival [8]. However, this is in agreement with several previous studies that found an extreme level of genomic instability associated with improved outcome compared to intermediate levels [26, 27].

A recent analysis of transcriptome subtypes in colorectal cancer generally questioned the existence of discrete subtypes, and proposed a continuum of transcriptomes instead [28]. Analogously dismissing the assumption of discrete subtypes for HGSOC rather warrants the notion of a spectrum for each of the four subtypes with transient boundaries between them. Such a subtype interpretation seems particularly plausible given also a recent study reporting a continuum of HGSOC genomes shaped by individual copy number signatures [18].

In agreement with the proposed model, our findings from single-cell subtyping imply a tumor at the transition from the differentiated to the immunoreactive spectrum. This was evident from the subtype calls on epithelial cells that were throughout at the border between differentiated and immunoreactive. The observation that subtype calls on single cells were typically less confident than on the corresponding bulk tumor likely results from a summarization effect. Small, but consistent expression changes towards the immunoreactive spectrum for individual cells thereby sum up across the bulk, which was more confidently assigned to the immunoreactive spectrum.

We also point out that the analysis of subtype association and subclonality of recurrent DNA alterations can be straightforward applied to other cancer types, as demonstrated for TCGA soft tissue sarcoma. However, concentrating the analysis on SCNAs is particularly suited for HGSOC and STS, both characterized by high levels of SCNAs and low levels of somatic mutations [19]. Using a combined approach that also takes into account somatic mutations is better suited for cancer types that are equally or especially driven by somatic mutations. Such an extension seems further warranted given that we found results from purity/ploidy-aware calling of DNA alterations to be highly consistent across platforms (whole-exome sequencing and SNP arrays) and computational methods (`ABSOLUTE` and `PureCN`).

We conclude that the previously proposed notion of four discrete subtypes does not realistically represent the genomic complexity of HGSOC. We propose a continuous subtype model in which HGS ovarian tumors evolve from a still largely intact genome (early DIF spectrum) towards a comprehensive loss of genome integrity (late PRO spectrum). In this model, stochastic and individually different genomic alterations from a constrained set of evolutionary moves give rise to increasing genomic instability and subclonal expansion (intermediate IMR/MES spectrum) that ultimately converge in the late PRO spectrum. This provides ready explanation for ambiguity in HGSOC subtype classification, which we found not only to be present on the cellular level, but in the instance analyzed to also exceed classification ambiguity on the bulk tumor. With the anticipated availability of more single-cell data for HGSOC in the near future, further confirmation of this observation is warranted and should particularly target tumors at the critical IMR/MES transition.

Methods

Statistical analysis was carried out in `R` [29] using packages of the `Bioconductor` repository [30].

Subtype assocation of SCNAs

Regions of recurrent focal CN amplification and deletion as detected with `GISTIC2` [13] for TCGA HGS ovarian tumors were obtained from the latest run of the TCGA Firehose pipeline (2016-01-28). The regions were classified depending on their type (deletion / amplification) for each tumor by `GISTIC2` as either normal (0), loss / gain of a single copy (1), or loss / gain of two or more copies (2). Results from transcriptome clustering using consensus non-negative matrix factorization (CNMF), which assigned each tumor to one of the four reported subtypes [2], were also retrieved from the 2016-01-28 Firehose run. Association of the obtained focal `GISTIC2` regions with the four subtypes was tested by χ^2 test with $df = 6$. Multiple testing correction was carried out using the method from Benjamini and Hochberg [31] with an FDR cutoff of 0.1.

Subclonality of SCNAs

Results from the application of the `ABSOLUTE` algorithm [16] to TCGA HGS ovarian tumors genotyped by Affymetrix SNP 6.0 arrays were obtained

bioRxiv preprint doi: https://doi.org/10.1101/554394; this version posted February 19, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

from the Pan-Cancer Atlas aneuploidy study [32]. This included per-tumor estimates of purity, ploidy, subclonal genome fraction, and number of genome doublings as well as segmented absolute copy number calls classified as occurring clonal or subclonal for each tumor. `ABSOLUTE` calls were managed in the `R/Bioconductor` data class `RaggedExperiment`, which implements a general ragged array schema for genomic location data [33]. This facilitated summarization of `ABSOLUTE`'s subclonality calls in `GISTIC2` regions using the `qreduceAssay` function. A `GISTIC2` region was hereby called subclonal for one tumor at a time if it was overlapped by at least one subclonality call. `GISTIC2` peaks were extended by 500 kb up- and downstream to account for uncertainty of the peak calling heuristic.

Correlation of subtype association with subclonality

Using the χ^2 test statistic as the subtype association score S_A of an alteration (Figure 3A) and the fraction of tumors for which this alteration is subclonal as the subclonality score S_C (Figure 3B), Spearman's rank correlation was computed to assess the relationship between S_A and S_C. Statistical significance of the correlation was assessed using Spearman's rank correlation test. To account for non-independence of the occurrence of different SCNAs, we also carried out a permutation test, where we permuted the observed S_A values 1000 times and recalculated the correlation with the observed S_C values. A p-value was then obtained by calculating the fraction of permutations in which the correlation of the permuted setup exceeded the observed correlation.

Absolute copy number analysis of whole-exome sequencing data

Whole-exome sequencing data available for 324 TCGA HGS ovarian tumors was downloaded from the NCI's Genomic Data Commons (`https://gdc.cancer.gov`) and subjected to absolute copy number analysis with `PureCN` [34] as described elsewhere [22]. Comparison to `ABSOLUTE` results with respect to per-tumor purity and ploidy estimates as well as individual copy number calls was done for 277 intersecting samples.

Subtype classification on single cell sequencing data

Bulk and single-cell RNA-seq data for one fresh HGSOC specimen was obtained from the Supplementary Material in [23]. Subtypes were classified using the consensus classifier implemented in the `consensusOV` package [35]. The extended 800-gene signature for classification was derived by selecting the 200 most representative genes per TCGA subtype cluster based on differential expression as previously described [4]. TCGA bulk RNA-seq and microarray data were obtained using the `curatedTCGAData` package [36]. Downsampling of TCGA bulk RNA-seq data to match the coverage of the scRNA-seq data was carried out using the function `downsampleMatrix` of the `DropletUtils` package [37].

Research reproducibility

Results are reproducible using `R` and `Bioconductor`. Code is available from `GitHub` (`https://github.com/waldronlab/subtypeHeterogeneity`).

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Conception and design: LG, MRi, LW. Development of methodology: LG, GP, MB, MRi, TS, LW. Acquisition of data: LG, SO, BW, TS, LW. Analysis and interpretation of data: LG, SO, MRa, LS, GP, MB, LXQ, MRi, TS, LW. Writing of the manuscript: LG, LW. Administrative, technical, or material support: BW, MM, TS. Study supervision: LW. All authors reviewed, discussed, and approved the final version of the manuscript.

Acknowledgements

The authors thank Andrew Cherniak for providing `ABSOLUTE` copy number data including purity and ploidy estimates for tumors from The Cancer Genome Atlas. LG was supported by a research fellowship from the German Research Foundation (GE3023/1-1). GP was supported by grant 5P30CA006516-53, and LW by grants 1R03CA191447-01A1 and U24CA180996 from the National Cancer Institute of the National Institutes of Health.

Author details

[1]Graduate School of Public Health and Health Policy, City University of New York, 55 W 125th St, New York, NY 10027, USA. [2]Institute for Implementation Science and Population Health, City University of New York, 55 W 125th St, New York, NY 10027, USA. [3]Roswell Park Cancer Institute, 665 Elm St, Buffalo, NY 14203, USA. [4]Department of Obstetrics, Gynecology and Women's Health, University of Minnesota, 420 Delaware St SE, Minneapolis, MN 55455, USA. [5]University of Minnesota Masonic Cancer Center, 420 Delaware Street SE, Minneapolis, MN 55455, USA. [6]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard Medical School, 450 Brookline Avenue, Boston, MA 02215, USA. [7]University of Alabama Comprehensive Cancer Center, 1824 6th Avenue South, Birmingham, AL 35233, USA. [8]Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY 10065, USA. [9]Novartis Institutes for BioMedical Research, 250 Massachusetts Ave, Cambridge, MA 02139, USA.

References

1. Tothill, R.W., Tinker, A.V., George, J., *et al.*: Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clin Cancer Res **14(16)**, 5198–208 (2008)
2. The Cancer Genome Atlas Research Network: Integrated genomic analyses of ovarian carcinoma. Nature **474(7353)**, 609–15 (2011)
3. Helland, A., Anglesio, M.S., George, J., *et al.*: Deregulation of MYCN, LIN28B and LET7 in a molecular subtype of aggressive high-grade serous ovarian cancers. PLoS One **6(4)**, 18064 (2011)
4. Verhaak, R.G., Tamayo, P., Yang, J.Y., *et al.*: Prognostically relevant gene signatures of highgrade serous ovarian carcinoma. J Clin Invest **123(1)**, 517–25 (2013)
5. Konecny, G.E., Wang, C., Hamidi, H., et al.: Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. J Natl Cancer Inst **106(10)** (2014)
6. Waldron, L., Haibe-Kains, B., Culhane, A.C., et al.: Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. J Natl Cancer Inst **106(5)** (2014)
7. Waldron, L., Riester, M., Birrer, M.: Molecular subtypes of high-grade serous ovarian cancer: the holy grail? J Natl Cancer Inst **106(10)** (2014)

bioRxiv preprint doi: https://doi.org/10.1101/554394; this version posted February 19, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

8. Chen, G.M., Kannan, L., Geistlinger, L., *et al.*: Consensus on molecular subtypes of high-grade serous ovarian carcinoma. Clin Cancer Res (2018). doi:10.1101/1078-0432.CCR-18-0784
9. Lohr, J.G., Stojanov, P., Carter, S.L., *et al.*: Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. Cancer Cell **25(1)**, 91–101 (2014)
10. Beroukhim, R., Mermel, C.H., Porter, D., *et al.*: The landscape of somatic copy-number alteration across human cancers. Nature **463(7283)**, 899–905 (2010)
11. Zack, T.I., Schumacher, S.E., Carter, S.L., *et al.*: Pan-cancer patterns of somatic copy number alteration. Nat Genet **45(10)**, 1134–40 (2013)
12. The Cancer Genome Atlas Research Network: Comprehensive molecular portraits of human breast tumours. Nature **490(7418)**, 61–70 (2012)
13. Mermel, C.H., Schumacher, S.E., B, H., *et al.*: GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol **12(4)**, 41 (2011)
14. Solimini, N.L., Xu, Q., Mermel, C.H., *et al.*: Recurrent hemizygous deletions in cancers may optimize proliferative potential. Science **337(6090)**, 104–9 (2012)
15. Guichard, C., Amaddeo, G., Imbeaud, S., *et al.*: Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet **44(6)**, 694–8 (2012)
16. Carter, S.L., Cibulskis, K., Helman, E., *et al.*: Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol **30(5)**, 413–21 (2012)
17. Ganzfried, B.F., Riester, M., Haibe-Kains, B., *et al.*: curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. Database **2013**, 013 (2013)
18. Macintyre, G., Goranova, T.E., De Silva, D., *et al.*: Copy number signatures and mutational processes in ovarian carcinoma. Nat Genet **50(9)**, 1262–70 (2018)
19. The Cancer Genome Atlas Research Network: Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. Cell **171(4)**, 950–65 (2017)
20. Widemann, B.C., Italiano, A.: Biology and management of undifferentiated pleomorphic sarcoma, myxofibrosarcoma, and malignant peripheral nerve sheath tumors: State of the art and perspectives. J Clin Oncol **36(2)**, 160–7 (2018)
21. Klein, M.E., Dickson, M.A., Antonescu, C., *et al.*: PDLIM7 and CDH18 regulate the turnover of MDM2 during CDK4/6 inhibitor therapy-induced senescence. Oncogene **37**, 5066–78 (2018)
22. Oh, S., Geistlinger, L., Ramos, M., *et al.*: Reliable analysis of clinical tumor-only whole exome sequencing data. bioRxiv (2019). doi:10.1101/552711
23. Winterhoff, B.J., Maile, M., Mitra, A.K., *et al.*: Single cell sequencing reveals heterogeneity within ovarian cancer epithelium and cancer associated stromal cells. Gynecol Oncol **144(3)**, 598–606 (2017)
24. Zhang, Q., Wang, C., Cliby, W.: Cancer-associated stroma significantly contributes to the mesenchymal subtype signature of serous ovarian cancer. Gynecol Oncol (2018)
25. Schwarz, R.F., Ng, C.K., Cooke, S.L., *et al.*: Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. PLoS Med **12(2)**, 1001789 (2015)
26. Salomon-Perzynski, A., Salomon-Perzynska, M., Michalski, B., Skrzypulec-Plinta, V.: High-grade serous ovarian cancer: the clone wars. Arch Gynecol Obstet **295(3)**, 569–76 (2017)
27. McGranahan, N., Swanton, C.: Clonal heterogeneity and tumor evolution: past, present, and the future. Cell **168(4)**, 613–28 (2017)
28. Ma, S., Ogino, S., Parsana, P., *et al.*: Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis. Genome Biol **19(1)**, 142 (2018)
29. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018). R Foundation for Statistical Computing. https://www.R-project.org
30. Huber, W., Carey, V.J., Gentleman, R., *et al.*: Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods **12(2)**, 115–21 (2015)
31. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc **57(1)**, 289–300 (1995)
32. Taylor, A.M., Shih, J., Ha, G., *et al.*: Genomic and functional approaches to understanding cancer aneuploidy. Cancer Cell **33(4)**, 676–89 (2018)
33. Ramos, M., Morgan, M.: RaggedExperiment: Representation of Sparse Experiments and Assays Across Samples. (2017). doi:10.18129/B9.bioc.RaggedExperiment. `http://bioconductor.org/packages/RaggedExperiment`
34. Riester, M., Singh, A.P., Brannon, A.R., *et al.*: PureCN: copy number calling and SNV classification using targeted short read sequencing. Source Code Biol Med **11**, 13 (2016)
35. Chen, G.M., Kannan, L., Geistlinger, L., et al.: consensusOV: Gene Expression-based Subtype Classification for High-grade Serous Ovarian Cancer. (2017). doi:10.18129/B9.bioc.consensusOV. `http://bioconductor.org/packages/consensusOV`
36. Ramos, M., Schiffer, L., Waldron, L., et al.: Curated Data from The Cancer Genome Atlas (TCGA) as MultiAssayExperiment Objects. (2017). doi:10.18129/B9.bioc.curatedTCGAData. `http://bioconductor.org/packages/curatedTCGAData`
37. Lun, A., Griffiths, J., McCarthy, D.: Utilities for Handling Single-cell Droplet Data. (2018). doi:10.18129/B9.bioc.DropletUtils. `http://bioconductor.org/packages/DropletUtils`

Additional Files

Additional file 1 — Supplementary Figures
PDF document containing Supplementary Figures **??–??**

Additional file 2 — Analysis vignette
HTML document containing literate analysis output.

bioRxiv preprint doi: https://doi.org/10.1101/554394; this version posted February 19, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

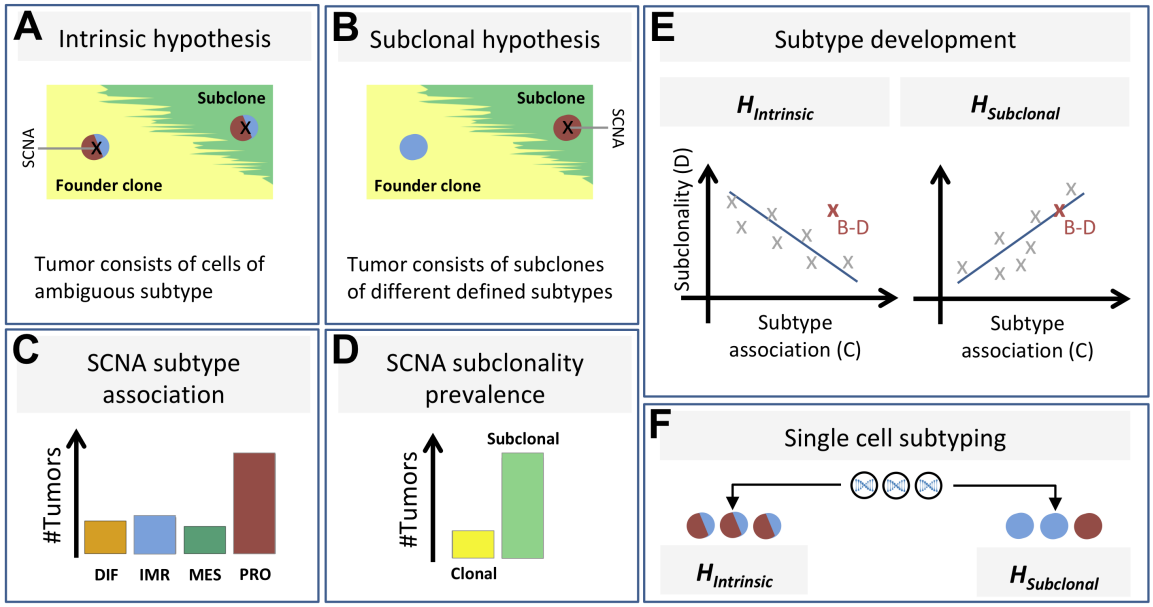

Figure 1 Study setup. Our study aims to distinguish between two possible hypotheses explaining why gene expression-based HGSOC subtypes are ambiguous. The intrinsic hypothesis **(A)** is that tumor cells display ambiguous expression patterns consisting of two or more subtype expression patterns. The subclonal hypothesis **(B)** is that a tumor contains multiple clones, with each clone displaying a consistent, yet distinct subtype expression pattern. To distinguish between these two hypotheses, we analyze recurrent SCNAs across many tumors and determine for each SCNA whether it occurs disproportionately often in tumors of a specific subtype **(C)**, and whether it occurs in the founder clone or a subclone **(D)**. The bar charts in **(C)** and **(D)** show here a particular SCNA associated with the proliferative subtype, occurring predominantly subclonally. If the subclonal hypothesis were true, there should be a positive correlation between SCNA subtype association and SCNA subclonality prevalence, while the intrinsic hypothesis predicts a negative correlation **(E)**. For example, the SCNA depicted in **(B-D)** (high subtype association and high subclonality) is more consistent with the subclonal hypothesis than with the intrinsic hypothesis (red X in **E**). However, only a trend across many recurrent SCNAs is considered evidence for either hypothesis. Analysis of single cell gene expression patterns **(F)** should also distinguish between the two hypotheses.

bioRxiv preprint doi: https://doi.org/10.1101/554394; this version posted February 19, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

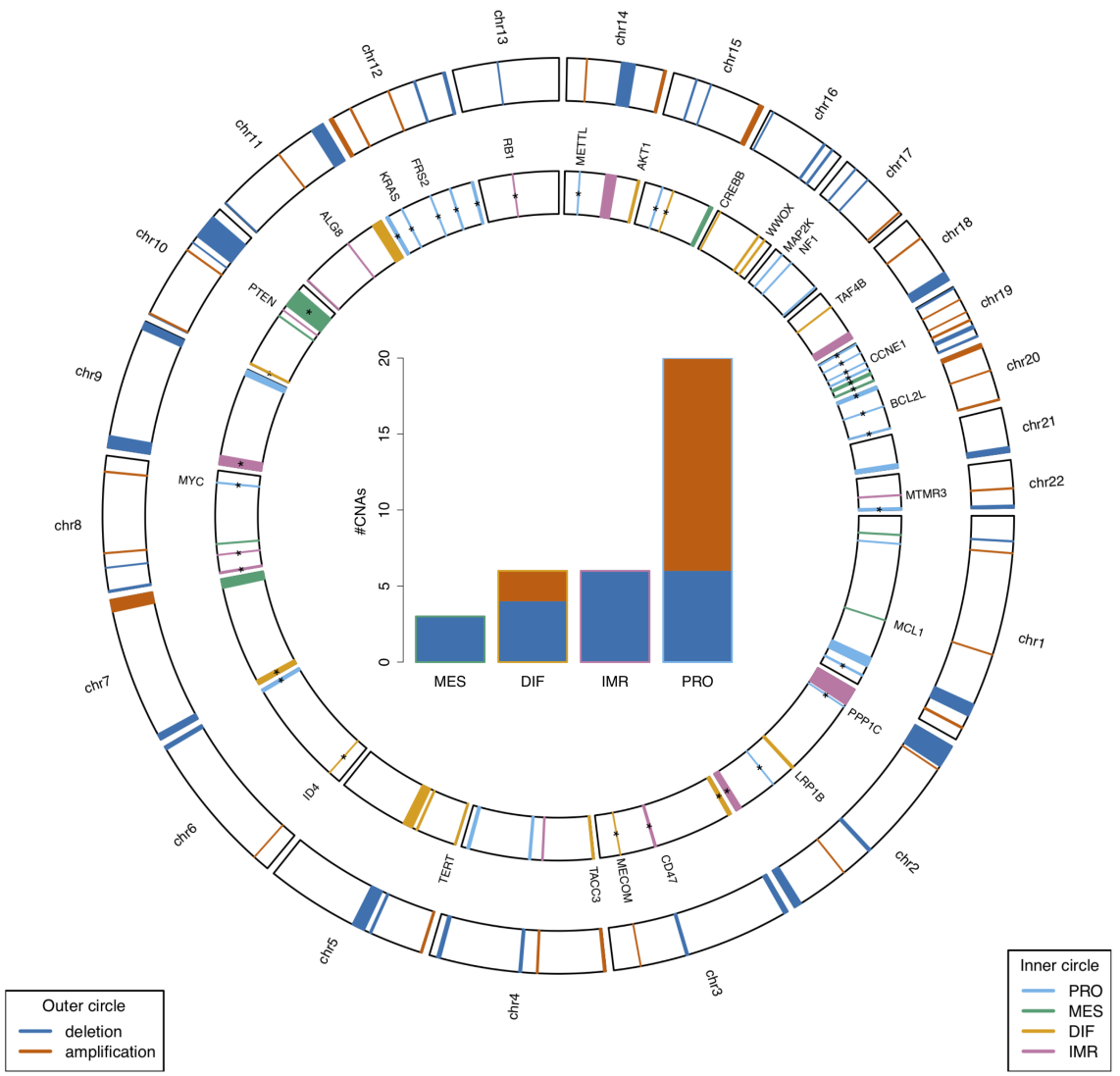

Figure 2 Genomic distribution of subtype-associated SCNAs. The circle on the outside shows the genomic location of focal CN amplifications (red) and deletions (blue) as detected with GISTIC2 [13] in TCGA HGS ovarian tumors. In the inner circle, the detected SCNAs are colored according to subtype association (blue: proliferative, green: mesenchymal, orange: differentiated, violet: immunoreactive). A star indicates significant association (FDR cutoff of 0.1, χ^2 test, Supplementary Figure **??**). For example, the MYC-containing amplification on chromosome 8 is significantly associated with the proliferative subtype as previously reported [2]. The barplot in the center shows for each subtype (x-axis) the number of significantly associated SCNAs (FDR < 0.1, y-axis) classified as deletion (blue) or amplification (red). Associations with the proliferative subtype are significantly overrepresented among the subtype-associated regions (20 out of 35, $p = 0.007$, Fisher's exact test).

bioRxiv preprint doi: https://doi.org/10.1101/554394; this version posted February 19, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

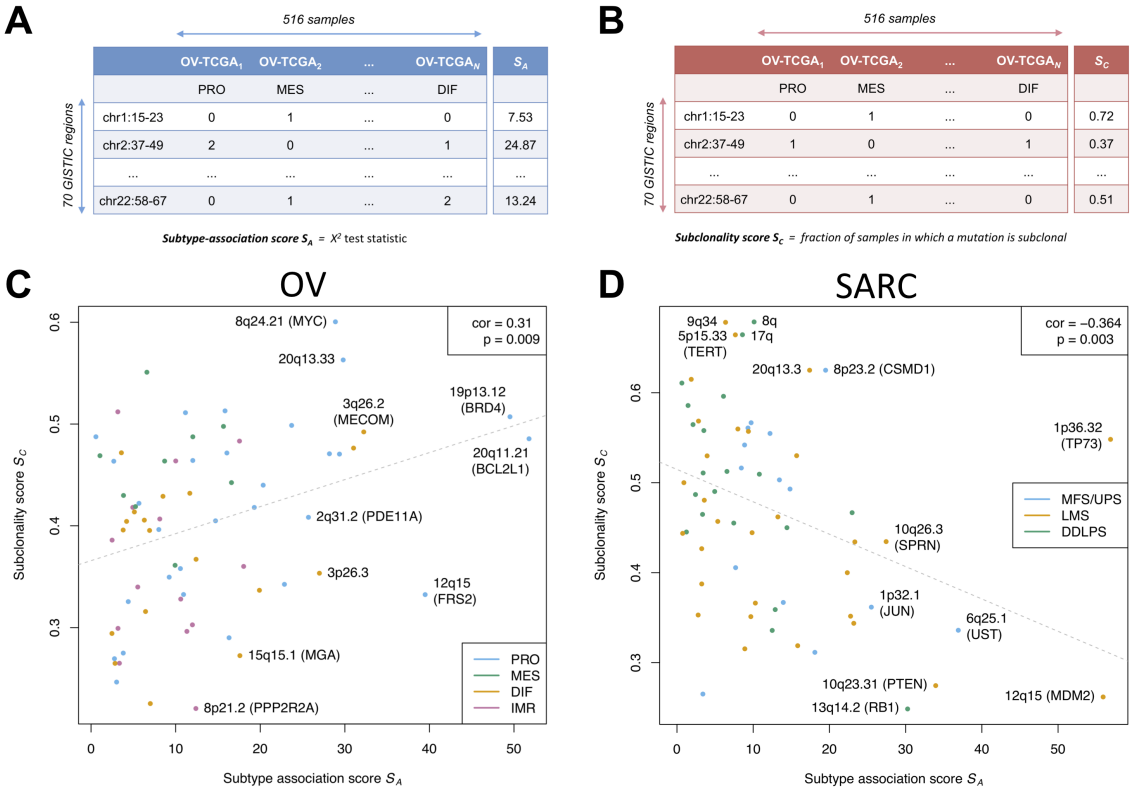

Figure 3 Correlation between subtype association and subclonality. The table in (**A**) illustrates the computation of the subtype-association score S_A. Rows of the table correspond to the focal SCNAs detected with `GISTIC2` (Figure 2). Columns correspond to the 516 TCGA HGS ovarian tumors, each one assigned to one of the 4 subtypes by transcriptome clustering (2nd row). The cells of the table indicate whether the region is of normal state (0), or contains a single (1) or higher (2) copy gain / loss, depending on whether the SCNA is rendered an amplification or deletion. The computation of the subclonality score S_C is analogously illustrated in (**B**). Here, the table cells indicate whether an SCNA is called as subclonal (1) or not (0) by `ABSOLUTE` [16]. The scatter plot in (**C**) depicts the correlation between S_A and S_C, showing a significant positive correlation of 0.31 with a p-value of 0.009 (Spearman's correlation test). The color of the dots corresponds to subtype association as in Figure 2. Correlation when stratifying tumors by purity is shown in Supplementary Figure **??**. Repeating the analysis outlined in (**A-C**) for TCGA soft tissue sarcomas (64 focal SCNAs, 259 samples) results in a significant negative correlation of -0.36 with a p-value of 0.003 as depicted in (**D**).

bioRxiv preprint doi: https://doi.org/10.1101/554394; this version posted February 19, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

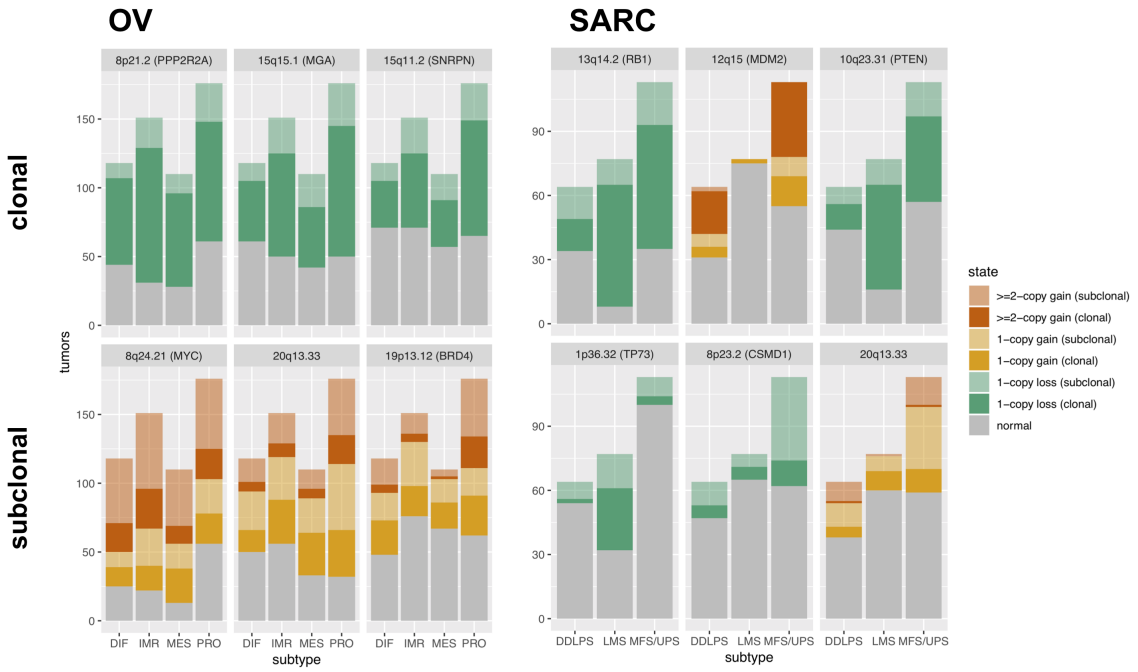

Figure 4 Predominantly clonal or subclonal copy number alterations. The barplots illustrate individual subtype-associated GISTIC2 regions from Figure 3C,D that occur predominantly clonal (top panel, solid color) or subclonal (bottom panel, transparent color) in TCGA HGSOC (left) or STS (right) cases. Each individual barplot displays the number of tumors (y-axis) of particular subtype (x-axis) that carry either a 1-copy loss (green), 1-copy gain (yellow), or $\geq$2-copy gain (red) in the region indicated at the top of each plot.

bioRxiv preprint doi: https://doi.org/10.1101/554394; this version posted February 19, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

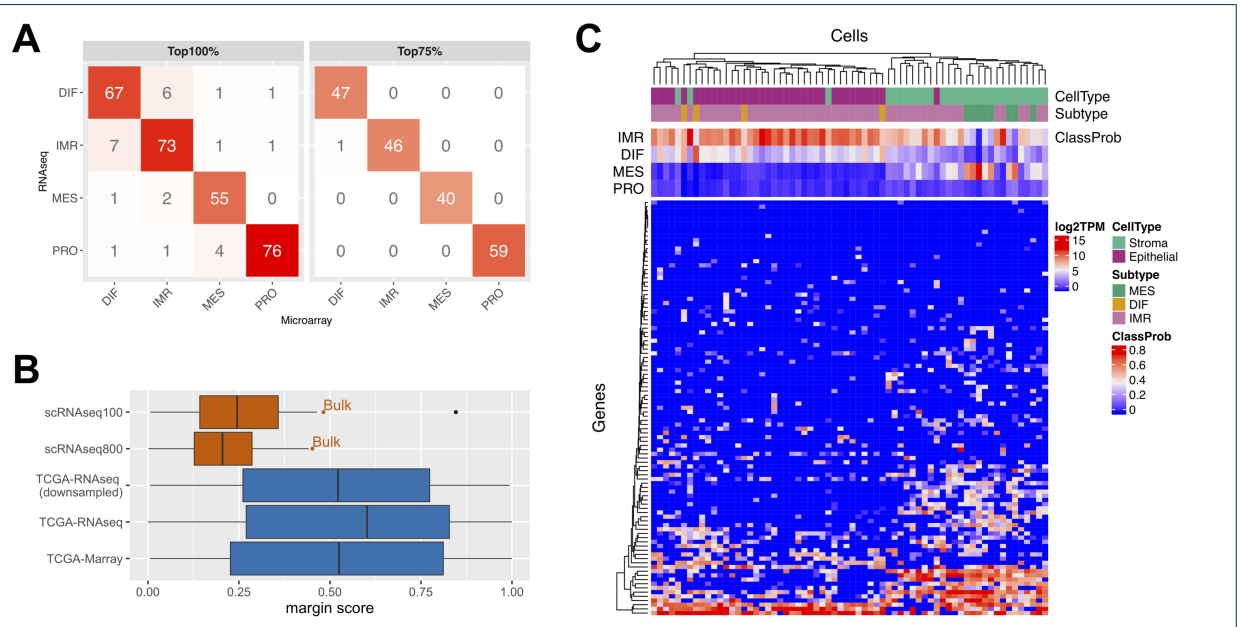

Figure 5 Single cell subtyping. (**A**) shows resulting subtype calls when applying the `consensusOV` [8] classifier to TCGA HGS ovarian tumors for which transcriptomic quantification was available via microarray (x-axis) and RNA-seq (y-axis). Concordance of subtype calls is shown for all tumors (top 100%, left panel) and the top 75% based on margin scores (right panel). The boxplots in (**B**) show margin score distributions of the subtype classification on 66 single cells (red) from Winterhoff *et al.* [23] as compared to the TCGA bulk classifications (blue) analyzed in (**A**). Single cell margins are shown when using the default 100-gene signature of the consensus classifier (*scRNAseq100*, detailed in (**C**)) and the extended 800-gene signature (*scRNAseq800*, detailed in Supplementary Figure **??**). The margin score of the bulk tumor is indicated in both cases by the red dot. The blue box named *TCGA-RNAseq (downsampled)* shows the margin score distribution when downsampling TCGA bulk RNA-seq data to match the coverage of the scRNA-seq data. The heatmap in (**C**) depicts log2 TPM expression values of the 100-gene signature used by the consensus classifier (rows) across 66 cells (columns). The annotation bars at the top show resulting subtype calls (*Subtype*), previous characterization of cells as epithelial or stromal by Winterhoff *et al.* [23] (*CellType*), and classification probabilities for each subtype (*ClassProb*).