nature research

Peer Review Information

Journal: Nature Genetics
Manuscript Title: Comprehensive Molecular Profiling of Multiple Myeloma Identifies Refined Copy Number and Expression Subtypes
Corresponding author name(s): Dr Jonathan (J) Keats, Dr Sagar Lonial

Editorial Notes:

Transferred manuscripts This document only contains reviewer comments, rebuttal and decision letters for versions considered at Nature Genetics.

Reviewer Comments & Decisions:

Decision Letter, initial version:

3rd Aug 2021

Dear Dr Keats,

Your Article, "Genomic Basis of Multiple Myeloma Subtypes from the MMRF CoMMpass Study" has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

You'll see that the reviewers are unanimously enthusiastic about the dataset that you have developed. Broadly, all three are supportive about the paper but issues have been raised about the overall novelty. To this, Reviewers #1 and #3 all suggest that a more integrated analysis will provide further biological insight and provide a more substantial advance. We agree that these analyses would strengthen the paper significantly and we would encourage you to perform them. All other reviewer comments should be addressed in full. Please do not hesitate to get in touch if you would like to discuss these issues further.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact

us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

Please use the link below to submit your revised manuscript and related files:

[redacted]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within 3-6 months. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Safia Danovi
Editor
Nature Genetics

Referee expertise:

Referee #1: Myeloma genomics

Referee #2: Plasma cell biology

Referee #3: Myeloma genomics

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Here Skerget et al present the findings of a large observational clinical study of newly diagnosed patients. Whole genome, exome, and transcriptome sequencing has been performed on up to 1143 patients from 84 clinical sites across the globe. The study itself is extremely important for the myeloma research community as the disease was left out of the cancer genome atlas (TCGA) and without it there would be no large genomic dataset to perform the important studies that have come out of it.

The MMRF have made the genomic and clinical data available through dbGAP with semi-annual updates. The scientific merit of the study is unprecedented. However, given that the data has been available online for several years it means that others have performed similar analyses in the past, making these results less novel.

The results presented here have been published in several different ways by other groups, and even by some of the co-authors listed on this manuscript. In 2018 Walker et al Blood (PMID:29884741) used this dataset in combination with the Myeloma XI sequencing dataset to identify the frequently mutated genes, the genomic driver events, copy number subtypes, and associations between them all. To some extent this is recapitulated through lines 119-182. There are some differences in the presentation of the data and some different genes are identified (e.g. CDC42BPB, MAGEC1) and differences in the copy number subgroups, but it is not clear why previously published copy number subgroups are better/worse that the ones presented here.

The analysis of chromosome 13 deletion is more novel (lines 134-145), with the suggestion of multiple genes on the chromosome suffering loss of function being PSPC1 and TGDS, although they are affected at much lower levels (1.5-1.9%) than previously identified genes RB1 and DIS3 (3.2-6.9%). A map of chromosome 13 and the events may help the reader understand why these new LOF events are just as important as the previously identified drivers that are dysregulated at higher frequencies.

Are these new events just passengers being carried along with the more frequent known driver events? There is also no mention of the mir15a/16-1 locus that has more recently been identified as a key driver of myelomagenesis (Chesi et al Blood Cancer Discovery; PMID 32954360). Was this locus not identified in the dataset at all?

The analysis of RNA subtypes (lines 183-259) has not been performed to my knowledge in RNA-seq data. However, many of the findings are identical to those previously found in gene expression array datasets (Zhan et al Blood 2006), and indeed the observed clusters are correlated back to these initial clusters. The novel findings are the subdivision of the CD2 subtype into CD2a and CD2b – although it is not clear what the difference is caused by or if this is important biologically or clinically.

The annotation of expression data with genomic features is more novel and helps explain some of the expression subgroups, but it would be more novel to integrate the expression and genomic data together to determine how these datasets combine to drive subgroups of myeloma. For example, in figure 2 five copy number groups are defined as hyperdiploid subtypes, defined by gain 3, 7, 15 or 11. But in Figure 3 the expression data cluster the same patients into hyperdiploidy with MYC abnormalities, TP53 or gain 15. So how do these 2 clusterings using different datatypes coalesce to identify what the important events are in myeloma? What should be used clinically to define myeloma patients – DNA level information or RNA level information? Are two methods better than one? Previously it has been very difficult to get genome and expression information on a large number of patient samples, but here this group is in a unique position to do that and truly define what the subgroups of myeloma should be based on integrative genomics.

The section on progression of disease subtype (lines 283-296) was novel until very recently, where a similar study was published by Boyle et al (BJH 2021).Here the authors have a similar finding to that published using serial samples from patients. The BJH paper had 147 patients with paired diagnosis/relapse samples and gene array data, whereas here the authors have 55 patients with RNA-seq data. Both groups find that patients can transition into the PR group and those that do have a worse outcome. The additional data form DNA studies here indicates a possible association with loss of negative regulators of cell cycle in patients who move the PR group.

Overall, this study is robust and excellently executed. There is little to comment on for the methods which have undergone refinement over the years to improve upon the unreliable cytogenetic/FISH data generated with world leading genomic analyses. However, I feel that there is little new information here that hasn't previously been published. Integration of the datatypes would be a key next step that has not been done yet.

Reviewer #2:
Remarks to the Author:
This manuscript describing the MMRF CoMMpass study, a large longitudinal tracking of newly diagnosed MM patients. The authors use a variety of techniques including whole genome, exome and RNA sequencing approaches to provide a wealth of data on MM diversity and progression. The study provides an excellent resource for the field and its findings can be rapidly incorporate into practice, particularly clinical trial design. They identified 8 clear patient clusters based on DNA analysis that appear to be well supported by the data. The transcriptome analysis appeared less easy to digest, with 12 clusters that don't fully match the genomic groupings. Finally, they identify a cohort of

patients that progress during the study period that have a very poor outcome.

Overall, this is a very useful dataset in terms of it scale and level of detail. The approach has the necessary power to cluster patients at high resolution and the provides many insights that need to be considered in future clinical trials in MM. While many aspects of the study overlap with previous much smaller cohorts, the scale of the undertaking as well as the use of previously untreated patients as a starting point is ideal for the aims of this project. I have only a few issues that should be considered before publication.

1. I have some questions about the low purity cluster which is 12% of samples. Were these preparations subject to the QC for purity in the bead isolation step? Is the low purity cluster entirely driven by the contaminating cells? Is this being the case what is the rationale for keeping these samples in the database? If the most contaminated outliers are removed using the techniques shown in Supp 3.12 does the cluster remain distinct?
2. Data accessibility. This paper is clearly a major resource for the MM field, yet at times it is difficult to access the specific data. For example, Figure 1C lists GOF events that were detected in 27 genes, where in most instances the GOF is reported to be due to a mutation. Yet there is no clear link to get from this Figure (or the results text section discussing this data) to get to the data to see what the mutations actually are. This is the type of data that we would find of interest and should be as easily accessible as possible to be of maximal use. This data may be available at the MMRF website, however, if so this type of information should be earmarked throughout the results so that those interested can track it down. Assuming that the MMRF site has the data (I didn't register for a login so cannot check), then a more detailed description of what is available and where it is would be useful in the "data availability" section.
3. The authors should provide some analysis of the genes driving the interesting CN2a vs b cluster split. The current data in Suppl 3 shows only the similarities between CN2a and b, such as CCND1 translocation, PAX5 and CD20 expression etc that does not help support the argument for the distinctiveness of the 2 clusters.

Reviewer #3:
Remarks to the Author:
The authors present an outstanding analysis of the CoMMpass dataset that they have generated and made publicly available. This dataset has proven to be an invaluable tool and the subject of multiple previous analyses. Nonetheless, the authors provide many important new insights with their analysis. These include
1 a comprehensive listing of genes with LOF and GOF using an integrated analysis of all of the data (WXS, WGS, RNAseq) they have generated. This is much more powerful than simply listing the frequency of SNV as is normally done.
2 A stellar analysis of copy number, with an unsupervised analysis identifying 8 subgroups. Although similar efforts have been published, the rigor of their analysis suggests to me a definitive analysis that is unlikely to change over time.
3 The unsupervised gene expression analysis reproduces features that have been noted in two previous classifications (UAMS and HOVON). This is a major improvement on those previous efforts, with much better correlation of the groups identified to the underlying genetics. This classification balances primary genetic events (IgH Tx, HRD), with shared mechanisms of disease progression (1q, PR, MYC). The new names given for subgroups are an improvement over previous names suggested

by UAMS and HOVON. There are also novel and important findings: the distinction between CD2a and CD2b is likely to be important. Although survival is similar, CD2b has more frequent 1q+, and higher PI and seems likely to represent a more aggressive subgroup. It is interesting that the Low Purity subgroup is predominantly HRD, likely representing a different disease biology, and not just poor sample quality.
4 A post=hoc integration of GOF/LOF with GEP subgroups is very useful and provides new insights (depletion of NRAS in MS and 1q gain, enrichment of FAM46C in ++15, IRF4 in CD2a and CD2b but not CD1, RB1 and MAX in PR)
5 Perhaps most useful is the analysis of serial samples that shows the PR subgroup as a common phenotype for progression

Although they have done excellent unsupervised analyses separately of CN and GEP, and superimposed selected features from other modalities, missing is an integrated unsupervised analysis starting from all WXS, WGS, CN, SV, GEP data. I think this was a wise decision, as it very artificial to assign weights to the various genetic measurements.

In summary it is an incredibly useful dataset, with an accompanying analysis that provides multiple important insights.

Minor points:

Median OS 74 month. There can be little confidence in this number. They should provide confidence intervals (or even better, say it is too early to determine).

Fig 1b. They should mention the major conclusion is that the majority of patients with del13 do not have complete LOF of a gene on chr13. One explanation for this is that in most patients the selective pressure for the loss of 13 is driven by loss of a single copy of one or more genes on chr13, as has been shown recently for mir15/16a (Chesi, Blood Cancer Discovery).

Fig 1c. Definition of GOF. Why is WHSC1 a GOF in almost all MS subgroup, but MAF and CCND1 in only a small fraction of the MAF and CD subgroups respectively? It suggests their parameters for over-expression with structural variation may need to be tweaked (or explained better)

Fig 3a. the figure shows MYC STR, but there is no mention of what this is in the text or figure legend. Presumably it is MYC structural variation (frequently abbreviated SV). This should be added to figure legend. They should add to methods how they define MYC STR. Table S7 provides the subtype for each patient in this figure. It would be very helpful to add the results for each of the columns from this figure to table S7 (e.g., HRD, +1q,…,PI, CD20). THIs would allow the reader to understand how each patient was classified for each abnormality, and to reproduce the figure. Looking at this figure it appears MYC STR is the most common structural abnormality in MM. It is surprising they do not mention the frequency in the text or tables. At the least they should add the frequency to Table 1 (which currently gives only the frequency of MYC – immunoglobulin translocations.)

Fig 3d,e. Checkpoint and CD expression in PR is not that interesting.

Author Rebuttal to Initial comments

Reviewer #1

Remarks to the Author:
Here Skerget et al present the findings of a large observational clinical study of newly diagnosed patients. Whole genome, exome, and transcriptome sequencing has been performed on up to 1143 patients from 84 clinical sites across the globe. The study itself is extremely important for the myeloma research community as the disease was left out of the cancer genome atlas (TCGA) and without it there would be no large genomic dataset to perform the important studies that have come out of it.

The MMRF have made the genomic and clinical data available through dbGAP with semi-annual updates. The scientific merit of the study is unprecedented. However, given that the data has been available online for several years it means that others have performed similar analyses in the past, making these results less novel.

As the reviewer points out the MMRF, a patient driven non-profit organization, created this unprecedented study. They chose a unique approach to making data available to the community with no specific embargo on pan-data set publications (as with TCGA and ICGC). The molecular data and clinical data are updated every six months, first released to supporting pharmaceutical companies, then six months later to participating clinical sites (who have a six month embargo on the data), and then one year after initial release the data is made public via the MMRF research gateway (https://research.themmrf.org/), dbgap (phs000748), and the GDC. As the team creating this valuable dataset which is widely championed as the defacto dataset in the field, we are proud of how much it has helped advance the field. However, as a consequence of early data release, observations we shared in meetings were often published by others using our own data. This report is the core paper from the study; outcomes are collected uniformly from the one protocol and sample data generated at a single center and we bring together several new observations including, importantly, a description of different mechanisms that mediate a very high risk phenotype at diagnosis that many patients transition to eventually at progression. This is the most comprehensive and updated version of this genomic and clinical data available and thus has a level of maturity and robustness not represented in other papers from non-CoMMpass investigators It is our hope that the data presented in this manuscript and the substantial reduction in sequencing costs coming in 2023-2024 will drive the adoption of whole genome sequencing in the field to improve patient prognostication and treatment selection.

1. *The results presented here have been published in several different ways by other groups, and even by some of the co-authors listed on this manuscript. In 2018 Walker*

et al Blood (PMID:29884741) used this dataset in combination with the Myeloma XI sequencing dataset to identify the frequently mutated genes, the genomic driver events, copy number subtypes, and associations between them all. To some extent this is recapitulated through lines 119-182. There are some differences in the presentation of the data and some different genes are identified (e.g. CDC42BPB, MAGEC1) and differences in the copy number subgroups, but it is not clear why previously published copy number subgroups are better/worse that the ones presented here.

Response:

The copy number clustering presented in this manuscript utilizes a distinct approach. First it uses the whole genome sequencing data from this dataset not the whole exome data used by Walker et al. that is approximately 3x more noisy in our hands. In the CoMMpass design the low-pass WGS was always the intended data source for copy number measurements, not the exomes. Second, the inputs into the clustering methods follow very different approaches that both have merit in our opinion. In the case of Walker et al. they first identified 39 focal regions associated with recurrent copy number alterations or markers for the classic trisomic chromosomes. Conversely, we chose to not weight specific regions of the genome over others and chose to weighted each 100kb bin of the genome equally. This difference in approach identifies very different groups and in our case two groups have effects on clinical outcome that could readily be used in clinical practice. In particular it identifies a group of hyperdiploid patients, HRD$^{1q+,\ diploid\ 11,\ -13}$, that have outcomes similar to ISS stage 3 patients. This approach also provides equal weight to chromosomes that were not even included in the Walker et al model, chromosomes 2, 10, 18, 20, and 22. Finally through this approach we identified 5 different hyperdiploid groups compared to the two identified by Walker et al. Less the aforementioned poor outcome hyperdiploid subtype, this analysis and the RNA subtype associations with DNA copy number have highlighted how common tetrasomy is on chromosome 15. Moreover, this analysis highlights how samples with gains on all of the classic hyperdiploid chromosomes (chr 3, 5, 7, 9, 11, 15, 19, and 21) have the best outcomes of all myeloma patients. Conversely two intermediate outcome groups of hyperdiploid are characterized by a diploid status of chromosome 7 or chromosomes 3 and 7. We suspect this latter observation is significant towards our understanding of hyperdiploid biology as it suggests a dependency where patients without trisomy 3 can not have trisomy 7. Chromosome 13 analyses are another distinguishing feature to this study (see below). Having access to all of the datasets from Commpass with updated clinical data allows us to make these unique observations not seen with other preliminary, interim data releases.

2. *The analysis of chromosome 13 deletion is more novel (lines 134-145), with the suggestion of multiple genes on the chromosome suffering loss of function being*

PSPC1 and TGDS, although they are affected at much lower levels (1.5-1.9%) than previously identified genes RB1 and DIS3 (3.2-6.9%). A map of chromosome 13 and the events may help the reader understand why these new LOF events are just as important as the previously identified drivers that are dysregulated at higher frequencies. Are these new events just passengers being carried along with the more frequent known driver events? There is also no mention of the mir15a/16-1 locus that has more recently been identified as a key driver of myelomagenesis (Chesi et al Blood Cancer Discovery; PMID 32954360). Was this locus not identified in the dataset at all?

Response:

We agree these observations are important to better understand the biology of chromosome 13 events which include known TSG, RB1 and several potential TSGs in myeloma. Figure 1B depicts LOF events on chromosome 13, with genes that are the target of recurrent loss of function. Similar to visualization with a map of chromosome 13, the genes are presented in chromosomal order, and contiguous genes are denoted. With this visualization, we were able to layer the events that occur in individual patients along the x-axis to visualize LOF events that are independent of RB1 and DIS3. As suggested we have added a clarifying sentence to the figure legend ("Each tick along the x-axis represents a patient with the corresponding event.").

Figure 1B shows a number of patients (located to the right of the end of the RB1 Complete Loss bar) with complete LOF events in other genes on chromosome 13 that occur in the absence of concomitant complete LOF events in either RB1 or DIS3. and thus, these genes likely represent novel drivers as they are not passenger events to loss of RB1 or DIS3.

Reviewer 1 (and 3's) comments on chr13 monosomy are helpful. The results in Figure 1B focus on chr13q genes with complete LOF (biallelic loss) in at least five patients in the baseline cohort. We did not identify any patients with complete LOF of mir15a/16-1, which is consistent with Chesi et al. who reported monoallelic deletion of mir15a/16-1 as a driver in myelomagenesis. These observations confirm and extend Chesi et al.: many CoMMpass patients had monoallelic loss of mir15a/16-1 due to monosomy 13 or, notably, rare interstitial deletions. Reviewer 3 noted that we did not identify complete LOF of a gene on chr13q in nearly half of patients with monosomy 13 and suggested this may be driven by monoallelic loss of one or more genes on chr13, as observed for mir15a/16-1. To further highlight this point and address these helpful comments, we added a sentence to the discussion and also added a reference for the suggested paper (Chesi, Blood Cancer Discovery) ("Notably, no complete loss event was identified in nearly half of patients (48.1%) with monosomy 13, highlighting that partial loss events resulting in haploinsufficiency, as observed in Mir15a/Mir16-1, may also drive this phenotype in myeloma").

3. *The analysis of RNA subtypes (lines 183-259) has not been performed to my*

knowledge in RNA-seq data. However, many of the findings are identical to those previously found in gene expression array datasets (Zhan et al Blood 2006), and indeed the observed clusters are correlated back to these initial clusters. The novel findings are the subdivision of the CD2 subtype into CD2a and CD2b – although it is not clear what the difference is caused by or if this is important biologically or clinically.

Response:

To minimize confusion, where appropriate, we have adopted the cluster names used by Zhan et al, which were also largely adopted by Broyl et al. We believe this highlights the reproducibility of all three studies and also limits confusion in the field about these entities. There are some important differences between the studies. For instance the low bone disease group from Zhan et al., which was not confirmed in Broyl et al. lines up with our 1q gain group and perhaps because of more advanced bone imaging there is no significant reduction of bone disease. The HY groups identified by both Zhan and Broyl were associated with hyperdiploid, and in our analysis you can clearly see that the genes used by this group identify our 4 different HRD groups (suppl fig 3.5 and 3.6), however, they are most associated with the HRD^{++15} and $HRD^{++15,MYC}$ subtypes we define and associated with increase incidence of tetrasomy 15 with or without a MYC structural abnormality. The PRL3 subtype identified by Broyl et al was most associated with one of our subtypes but the associated expression index was also associated with our 1q gain, MS and PR subtypes (suppl fig 3.6). As a consequence we looked to identify features that defined this group and identified frequent MYC structural events and a low NF-kB index only otherwise seen in the PR population (suppl fig 3.10) and hence we called the subtype $HRD^{MYC,low\ NFKB}$. This low NF-kB index highlights a very unique biological feature of this hyperdiploid subtype and the PR subtype. The NF-kB subtype identified by Broyl et al. was most closely associated with one of our subgroups but there was no clear association with a high NF-kB index (suppl fig 3.10), which lead us to identify the association of this subtype with high NINJ1 and low TP53 expression (suppl fig 3.11). We also correctly identify the myeloid population from Broyl et al, not as a distinct biological entity but a group of samples with non B lineage contamination. As noted we also subdivide the CD2 population from Zhan et al and Broyl et al. into two groups that we called CD2a and CD2b given the clear association (suppl fig 3.5 and 3.6). Biologically the CD2b subtype had a higher proliferative index and a trend to inferior PFS and OS outcomes. To further clarify these observations, the following sentences were added to the Results section: "Compared to CD2a, patients in the CD2b subtype had higher proliferative index scores ($p<0.005$) but there is not a significant difference in OS or time to second line therapy between these groups and unexpectedly, CD2a patients start second line therapy 8.4 months earlier." and the Methods section "The Mann-Whitney test was used to compare proliferation index scores for the CD2a versus CD2b subtypes.". Finally and importantly, we identified the genetic basis for the PR population which provides an opportunity for a subset of these patients to be identified with

comprehensive DNA based approaches like whole genome sequencing.

4. *The annotation of expression data with genomic features is more novel and helps explain some of the expression subgroups, but it would be more novel to integrate the expression and genomic data together to determine how these datasets combine to drive subgroups of myeloma. For example, in figure 2 five copy number groups are defined as hyperdiploid subtypes, defined by gain 3, 7, 15 or 11. But in Figure 3 the expression data cluster the same patients into hyperdiploidy with MYC abnormalities, TP53 or gain 15. So how do these 2 clusterings using different datatypes coalesce to identify what the important events are in myeloma? What should be used clinically to define myeloma patients – DNA level information or RNA level information? Are two methods better than one? Previously it has been very difficult to get genome and expression information on a large number of patient samples, but here this group is in a unique position to do that and truly define what the subgroups of myeloma should be based on integrative genomics.*

Response:

We agree that combining structural features with expression data shows important insights. And that this unique, large dataset generated in one center enables many combinatorial analyses. A primary goal was to facilitate precision medicine approaches that are both analytically and clinically feasible in multiple myeloma. As indicated by reviewer #3, assigning weights to the different data types introduces biases, and the high dynamic range of gene expression data versus copy number data leads to a highly RNA driven subtype. Moreover, a dual approach, with DNA and RNA sequencing would require groups to adopt two new assays into the clinical testing stream. With our sequencing approach (Figure 2) it is possible groups might focus on whole genome sequencing, there by getting mutation detection, the copy number subtypes, the robust detection of all clinically relevant translocations (not just the limited IgH events tracked by most FISH assays today) and they would be able to identify a subset of PR patients based on loss of function events associated with this subtype. This single assay strategy would also facilitate the identification of mutations in binding sites or the loss of Car-T or T-cell engager antigens (recently highlighed by Friedrich et al. Cancer Cell 2023) .

5. *The section on progression of disease subtype (lines 283-296) was novel until very recently, where a similar study was published by Boyle et al (BJH 2021).Here the authors have a similar finding to that published using serial samples from patients. The BJH paper had 147 patients with paired diagnosis/relapse samples and gene array data, whereas here the authors have 55 patients with RNA-seq data. Both groups find that patients can transition into the PR group and those that do have a worse outcome.*

The additional data form DNA studies here indicates a possible association with loss of negative regulators of cell cycle in patients who move the PR group.

Response:

We thank the reviewer for highlighting this important work by Boyle et al, who used Affymetrix data from one large center to show that tumors with various RNA expression subtypes can transition into the PR subtype at relapse. We added a reference to this paper in the discussion. This report, using current gene expression technology, further demonstrates the frequency and importance of this PR transition, which highlights the need to define PR tumor biology. This study adds matched whole genome and exome analysis to finally identify the diverse genetic basis of the PR population (primarily, RB1 and MAX complete loss at diagnosis and CDKN2C complete loss at progression), which was always one of the major complaints about the unsupervised approaches versus the supervised TC classification approaches. Given the poor outcome associated with PR and it's distribution across cytogenetic and FISH defined subtypes, this linkage of PR to DNA based events will facilitate the identification of some of these high risk patients in clinical tests that are limited to DNA based measurements, including whole genome sequencing.

6. *Overall, this study is robust and excellently executed. There is little to comment on for the methods which have undergone refinement over the years to improve upon the unreliable cytogenetic/FISH data generated with world leading genomic analyses. However, I feel that there is little new information here that hasn't previously been published. Integration of the datatypes would be a key next step that has not been done yet.*

Response:

We appreciate the reviewer's positive opinion on the quality of the MMRF CoMMpass dataset. While we disagree on the relative novelty, we appreciate recognition of our work to make all data available to the entire research community without holding it to first publish our own analyses. We also appreciate suggestions for additional analyses. We have, as a patient centered, foundation funded endeavor, focused first on what we see as the most practical and potentially clinically impactful aspects. While potentially biologically insightful, we have not attempted the suggested unsupervised analysis of the integrated genomic measurements. This is informed by our concern, also articulated by Reviewer 3, that there are too many confounding variables, and that the clinical translation of results that require multiple data types would be significant. To maximize potential clinical translation, we focused on identifying subtypes from either DNA or RNA based

measurements. The DNA subtypes highlight the importance of 1q gain and 13q loss in hyperdiploid patients. Where we extend on those studies is by having the comprehensive dataset to delineate biologically relevant relationships with these groups and thus refine groups even further. In the case of the PR subtype we highlight the diverse array of events that likely mediate the loss of G1/S cell cycle control in this population. We also integrated the LOF and GOF analyses to highlight those genes that are likely acting as tumor suppressor genes or oncogenes in myeloma (eg. depletion of NRAS in MS and 1q gain, enrichment of FAM46C in ++15, IRF4 in CD2a and CD2b but not CD1, RB1 and MAX in PR) through a diversity of mechanisms based on the comprehensive measurements made in this cohort.

Reviewer #2

Remarks to the Author:

This manuscript describing the MMRF CoMMpass study, a large longitudinal tracking of newly diagnosed MM patients. The authors use a variety of techniques including whole genome, exome and RNA sequencing approaches to provide a wealth of data on MM diversity and progression. The study provides an excellent resource for the field and its findings can be rapidly incorporate into practice, particularly clinical trial design. They identified 8 clear patient clusters based on DNA analysis that appear to be well supported by the data. The transcriptome analysis appeared less easy to digest, with 12 clusters that don't fully match the genomic groupings. Finally, they identify a cohort of patients that progress during the study period that have a very poor outcome.

Overall, this is a very useful dataset in terms of it scale and level of detail. The approach has the necessary power to cluster patients at high resolution and the provides many insights that need to be considered in future clinical trials in MM. While many aspects of the study overlap with previous much smaller cohorts, the scale of the undertaking as well as the use of previously untreated patients as a starting point is ideal for the aims of this project. I have only a few issues that should be considered before publication.

We thank reviewer #2 for their comments on the value of the dataset to the entire myeloma field and the potential for it to drive future clinical trials. From the outset the intention of this study was to improve patient outcomes through a better understanding of the subtypes of multiple myeloma and how those subtypes respond to the diversity of therapies a myeloma patient encounters over their disease course.

Comments:

1. *I have some questions about the low purity cluster which is 12% of samples. Were these preparations subject to the QC for purity in the bead isolation step?*

Response:

All samples were CD138 enriched and the purity of the enriched fraction was assessed by either flow cytometry or using an established slide-based assay. In all cases samples were only moved forward for molecular analysis if these assays indicated >80% tumor cell content (median 93.9%, IQR 91.4-95.5). Though methods vary at the three biorepositories no correlation with low purity and biorepository was observed.

2. *Is the low purity cluster entirely driven by the contaminating cells? Is this being the case what is the rationale for keeping these samples in the database? If the most contaminated outliers are removed using the techniques shown in Supp 3.12 does the cluster remain distinct?*

Response:

Based on the contamination and purity estimates presented in Supplementary Figure 3.12 the contaminating cells are a minor fraction of the cell population profiled in each sample. We elected to keep these samples in the dataset because other modalities included here, namely whole genome and whole exome DNA sequencing, can still yield very valuable results from samples with approximately 7-15% more non plasma cells, as our thresholds for making variant calls, copy number calls, or structural calls equates to 10%, 20%, and 30% tumor content, respectively. However, we do believe it is the small fraction of contaminating cells that drives the gene expression clustering of the low-purity group. In this consensus clustering approach, the low purity group separates out at K=4, while by comparison the MAF cluster, which is the most similar within group cluster, only appears at K=6. Likely the dramatic increase in expression of cell type defining genes from even a few non tumor cells can override the subtle expression signature of most tumor subtypes identified.

As suggested, we have tried removing the contaminated samples based on the non B-cell contamination index, however, the only effect is that some samples that were previously in definitive clusters now shift over to a low purity subtype. Since this low purity subtype appears at such an early K in the consensus clustering approach, we believe it is just this method simply identifies a subset of samples with the most contamination; upon removal of some samples those samples with the next most contamination are clustered together. As illustrated in Supplementary figure 5.1, patients can transition between a definitive subtype and low purity with time, which illustrates this is a sample intrinsic effect.

3. *Data accessibility. This paper is clearly a major resource for the MM field, yet at times it is difficult to access the specific data. For example, Figure 1C lists GOF events that were detected in 27 genes, where in most instances the GOF is reported to be due to a mutation. Yet there is no clear link to get from this Figure (or the results text section discussing this data) to get to the data to see what the mutations actually are. This is the type of data that we would find of interest and should be as easily accessible as possible to be of maximal use. This data may be available at the MMRF website, however, if so this type of information should be earmarked throughout the results so that those interested can track it down. Assuming that the MMRF site has the data (I didn't register for a login so cannot check), then a more detailed description of what is available and where it is would be useful in the "data availability" section.*

Response:

We appreciate the effort of reviewer #2 to dig into the supplemental tables to extract the value another investigator might need to follow-up these observations. In the case of the LOF and GOF analysis, this was executed internally in a non-relational database to facilitate the complex data integration tasks. Unfortunately, this is not something that can be made available as a simple supplementary table; yet we will add this to a revised MMRF researcher gateway.

Though we appreciate it takes some effort it is possible, to extract this information from the supplemental tables today. All LOF and GOF events identified in the cohort are summarized in Supplemental Table 6. For each gene for each analyzed patient, we assigned a functional status and also designated what type of event(s) were detected to contribute to the assigned functional status. Using the flags, you can cross reference the other provided supplemental tables to view specific information about each event (for example, specific mutation positions, copy number states, TPM estimates etc.). Below, we have presented some examples of LOF and GOF events and describe how to cross reference the events in the other provided supplemental tables. In addition, we are providing an expanded version of Supplemental_Table_3_Delly_Structural_Calls.xlsx where we add columns designating the genes directly upstream and downstream of detected structural events, which we hope will further facilitate cross-referencing of structural events.

LOF/GOF Cross Referencing Examples

Note: the data presented in the tables below are derived from one row of the referenced table. Select columns are presented for brevity.

Example 1

Sample MMRF_1167_1_BM had complete LOF of RB1 due to a one copy loss and mutation.

This information is contained in Supplemental_Table_6_Gene_Impact.xlsx

COLUMN NAME	ENTRY
SAMPLE	MMRF_1167_1_BM
GENE	RB1
ENSG	ENSG00000139687
EFFECT	COMPLETE_LOSS
OneCopyLoss	1
OneMutation	1
ImapctCodeDescription	DeletionandMutation

In order to determine the specific events, you can cross reference the provided supplemental tables.

For the one copy loss, reference Supplemental_Table_5A_CNA_PerGene_LowestSegment.xlsx

COLUMN NAME	ENTRY
Gene	ENSG00000139687

MMRF_1167_1_BM	-1.004

Based on the thresholds specified in the Loss-of-Function (LOF) section of the methods (paragraph 3), a log2 value of -1.004 corresponds to a single copy loss (heterozygous deletion).

For the mutation, reference Supplemental_Table_2_Mutations.xlsx

COLUMN NAME	ENTRY
sample	MMRF_1167_1_BM
start	49033892
end	49033892
GENE	RB1
REF	GAACATATC
ALT	G
EFFECT	frameshift_variant
TUMOR_ALT_FREQ	0.866
GENEID	ENSG00000139687
HGVS_P	p.Glu677fs

Taken together the one copy deletion and mutation were identified as the two events leading to a complete loss-of-function in RB1 for sample MMRF_1167_1_BM.

Example 2

Sample MMRF_2107_1_BM had GOF of MAP3K14 due to overexpression and a structural variant.

This information is contained in Supplemental_Table_6_Gene_Impact.xlsx

COLUMN NAME	ENTRY
SAMPLE	MMRF_2107_1_BM
GENE	MAP3K14
ENSG	ENSG00000006062
Overexpression	1

ImapctCodeDescription	OEWithSV

For the structural variant, reference Supplemental_Table_3_Delly_Structural_Calls.xlsx

COLUMN NAME	ENTRY
Specimen_ID	MMRF_2107_1_BM
SVTYPE	TRA
LGENISEC_ENSG	ENSG00000006062
LGENISEC_HUGO	MAP3K14

For expression, reference Supplemental_Table_4_Salmon_V7.2_Filtered_Gene_TPM_and_InFrame_Fusions.xlsx

COLUMN NAME	ENTRY
GENE_ID	ENSG00000006062
MMRF_2107_1_BM	345.463

4. *The authors should provide some analysis of the genes driving the interesting CN2a vs b cluster split. The current data in Suppl 3 shows only the similarities between CN2a*

and b, such as CCND1 translocation, PAX5 and CD20 expression etc that does not help support the argument for the distinctiveness of the 2 clusters.

Response:

The most striking difference between the CD2a and CD2b groups is the difference in proliferation index shown in Supplemental Figure 3.8. This dichotomy is also reflected in the median time to second line therapy between these groups. We have extensively evaluated a number of parameters commonly associated with t(11;14) for associations with these two group from N-terminal CCND1 mutations, translocation alleles harboring the CCND1 c870G risk allele, translocation breakpoint regions on 11q13, or copy number subtypes and were unable to identify a statistically significant difference that might explain the two phenotype and their associated differences in proliferation and time to second line therapy. We added the following sentence to the results section to highlight the primary difference "Compared to CD2a, patients in the CD2b subtype had higher proliferative index scores ($p<0.005$) but there is not a significant difference in OS or time to second line therapy between these groups and unexpectedly, CD2a patients start second line therapy 8.4 months earlier"

Reviewer #3

Remarks to the Author:

The authors present an outstanding analysis of the CoMMpass dataset that they have generated and made publicly available. This dataset has proven to be an invaluable tool and the subject of multiple previous analyses. Nonetheless, the authors provide many important new insights with their analysis. These include

1 a comprehensive listing of genes with LOF and GOF using an integrated analysis of all of the data (WXS, WGS, RNAseq) they have generated. This is much more powerful than simply listing the frequency of SNV as is normally done.

2 A stellar analysis of copy number, with an unsupervised analysis identifying 8 subgroups. Although similar efforts have been published, the rigor of their analysis suggests to me a definitive analysis that is unlikely to change over time.

3 The unsupervised gene expression analysis reproduces features that have been noted in two previous classifications (UAMS and HOVON). This is a major improvement on those previous efforts, with much better correlation of the groups identified to the underlying genetics. This classification balances primary genetic events (IgH Tx, HRD), with shared mechanisms of disease progression (1q, PR, MYC). The new names given for subgroups are an improvement over previous names suggested by UAMS and HOVON. There are also novel and important findings: the distinction between CD2a and

CD2b is likely to be important. Although survival is similar, CD2b has more frequent 1q+, and higher PI and seems likely to represent a more aggressive subgroup. It is interesting that the Low Purity subgroup is predominantly HRD, likely representing a different disease biology, and not just poor sample quality.
4 A post=hoc integration of GOF/LOF with GEP subgroups is very useful and provides new insights (depletion of NRAS in MS and 1q gain, enrichment of FAM46C in ++15, IRF4 in CD2a and CD2b but not CD1, RB1 and MAX in PR)
5 Perhaps most useful is the analysis of serial samples that shows the PR subgroup as a common phenotype for progression

Although they have done excellent unsupervised analyses separately of CN and GEP, and superimposed selected features from other modalities, missing is an integrated unsupervised analysis starting from all WXS, WGS, CN, SV, GEP data. I think this was a wise decision, as it very artificial to assign weights to the various genetic measurements.

In summary it is an incredibly useful dataset, with an accompanying analysis that provides multiple important insights.

We thank reviewer #3 for their comments and acknowledgement of this dataset's importance now and in the future. We agree that a fully integrated unsupervised analysis is difficult as each modality has substantial confounders. Instead, we focused on identifying subtypes that could be reproduced with a single data type, DNA or RNA, and leveraging the diverse data in this cohort to identify genes that are likely tumor suppressor genes or oncogenes in multiple myeloma.

Comments:

1. *Median OS 74 month. There can be little confidence in this number. They should provide confidence intervals (or even better, say it is too early to determine).*

Response:

We agree with the reviewer that the early clinical outcomes used in the first version (IA14) and by many studies using MMRF CoMMpass data were too preliminary. To address this, we initiated a major initiative with the clinical research organization helping to manage CoMMpass to ensure all patients that were not lost to follow-up were updated in the IA22 release, which is still embargoed from publication for anyone except the lead CoMMpass team. This update pushes the median overall survival out to 104 months. We have noted the lower confidence limit in the manuscript but unfortunately, we will most likely never know the upper limit even in the final IA24 release. The estimated median now exceeds the planned 8 year observation time of the study, which was extended by 3 years when it became apparent that the initial 5

year observation window was to short given the dramatic improvement in outcomes seen in the last 1-2 decades.

2. *Fig 1b. They should mention the major conclusion is that the majority of patients with del13 do not have complete LOF of a gene on chr13. One explanation for this is that in most patients the selective pressure for the loss of 13 is driven by loss of a single copy of one or more genes on chr13, as has been shown recently for mir15/16a (Chesi, Blood Cancer Discovery).*

Response:

We agree with the reviewer that this is an important point. We included this as a primary figure as we believe it is important to highlight that no one gene is the clear target of monosomy 13 myeloma and that other hypotheses such as the potential for multiple target genes and an effect of haploinsufficiency must be considered. As such we added a sentence to the discussion and also added a reference for the suggested paper (Chesi, Blood Cancer Discovery) ("Notably, no complete loss event was identified in nearly half of patients (48.1%) with monosomy 13, highlighting that partial loss events resulting in haploinsufficiency, as observed in Mir15a/Mir16-1, may also drive this phenotype in myeloma")

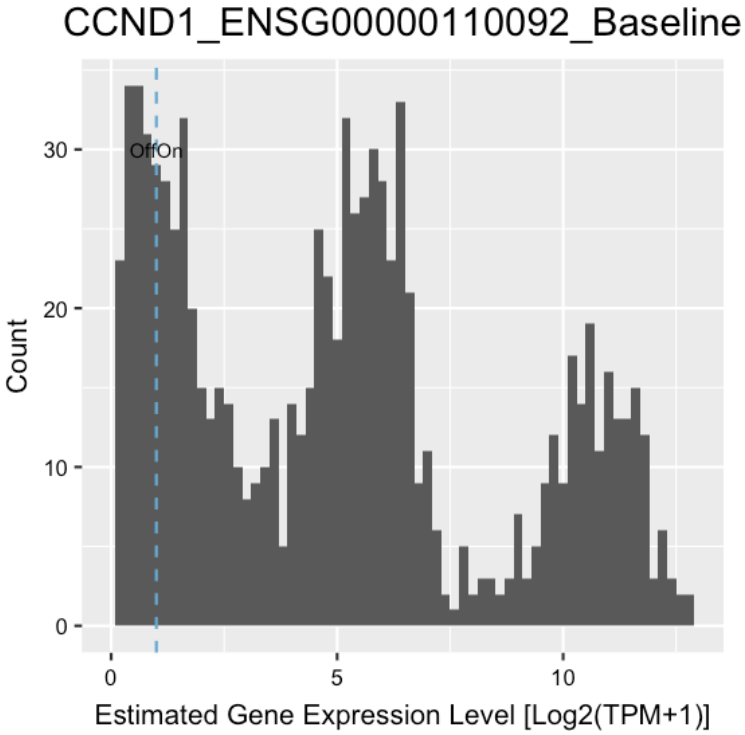
3. *Fig 1c. Definition of GOF. Why is WHSC1 a GOF in almost all MS subgroup, but MAF and CCND1 in only a small fraction of the MAF and CD subgroups respectively? It suggests their parameters for over-expression with structural variation may need to be tweaked (or explained better).*

Response:

This integrated analysis attempts to bring together a number of diverse data types into a single interpretation of each gene in each patient. For most of the data types the measurements are restricted to specific properties of the gene; mutation, copy number alterations, and fusions. In the case of gene overexpression we balanced false positive and false negative gene identification by requiring both overexpression and a proximal structural event. We identified genes with outlier expression (ie. those with TPM>=1, whose median absolute deviation exceeded the 75th percentile plus 1.5x the interquartile range) that were within or immediately adjacent to a structural breakpoint before the start or stop of the next coding gene on either side of the gene. These two constraints create limitations for the two events in question that when looked at based on WGS translocations in figure 3 line up as expected with most MF patients having a MAF, MAFA or MAFB translocation while most CD1/CD2a/CD2b have a

CCND1/CCND2/CCND3 translocation. In the case of CCND1, the trimodal distribution of expression (see figure below) limits the number of translocation positive cases with outlier expression. Then in some cases with breakpoints centromeric of MYEOV the breakpoint is past the next coding gene causing that event to not be flagged. This latter issue is very common in the case of MAF as the breakpoints can occur at distant sites centromeric of MAF within WWOX and there are a number of protein coding genes in between. We acknowledge that this is a limitation but it also helped to identify MAP3K14 as a GOF target, which we expect the reviewer would agree is valid and also flagged LETM1 as a GOF target in the MS patients which is likely a false positive. It was with this balance in mind that led us to select the parameters used for the analysis

To make the MAF issue more clear we updated a sentence in the methods to indicate the structural break must be within the gene in question or before the next coding gene is encountered on either side of the gene, “To limit false positive overexpression calls, we only set the ‘*Oe*’ bit when an accompanying DNA structural event was detected in the gene or immediately upstream or downstream before the start or stop of the next coding gene in chromosomal order.”



4. *Fig 3a. the figure shows MYC STR, but there is no mention of what this is in the text or figure legend. Presumably it is MYC structural variation (frequently abbreviated SV). This should be added to figure legend. They should add to methods how they define MYC STR. Table S7 provides the subtype for each patient in this figure. It would be very helpful to add the results for each of the columns from this figure to table S7 (e.g., HRD, +1q,…,PI, CD20). THIs would allow the reader to understand how each patient was classified for each abnormality, and to reproduce the figure. Looking at this figure it appears MYC STR is the most common structural abnormality in MM. It is surprising they do not mention the frequency in the text or tables. At the least they should add the frequency to Table 1 (which currently gives only the frequency of MYC – immunoglobulin translocations.)*

Response:

We would like to thank reviewer #3 for having caught this omission, that we did not define MYC STR in the text, figure legend, or methods. The importance of MYC in myeloma is unquestioned and we believe CoMMpass has helped to show it is more than just a late progression event, as such it is very important that the diversity of events that affect MYC are correctly documented. We have made a number of revisions to address this comment.

In the methods section, under the "Sequencing Data Analysis" subheading we added a sentence to explain how we derived the MYC STR flag ("The MYC STR flag was derived from Delly structural calls and indicates detection of either a (i) MYC translocation (Ig or non-Ig) or (ii) intrachromosomal deletion with a break end <3MB centromeric of, but not spanning, MYC."). In the Figure 3a legend we added a sentence defining the MYC STR flag ("The MYC STR flag indicates detection of a MYC translocation (Ig or non-Ig) or intrachromosomal deletion centromeric of MYC."). In Table 1, we added an entry listing the frequency of the MYC STR flag in the baseline cohort (251 patients, 29.5%), as well as a footnote explaining the MYC STR flag "*MYC translocation (Ig or non-Ig) or intrachromosomal deletion proximal to MYC". This now becomes footnote 2(**).
The RNA subtype of each patient at baseline and the data for each column as depicted in Figure 3a is listed in Table S1 which would allow the reader to cross reference the patient, RNA subtype, and molecular features. To avoid duplicating the data provided in supplemental tables, we did not add the data to Table S7 as requested.

5. *Fig 3d,e. Checkpoint and CD expression in PR is not that interesting.*

Response:

Given the very poor prognosis of the PR population in this cohort we believe it is important to highlight the potential directions for high-risk clinical trials in the future. Though there are limited differences, the statistically significant expression differences observed suggest common Car-T or T-cell engager targets are better targets in this high-risk population. We believe this adds value to the community and helps to justify these patients are not excluded from clinical trials testing these agents.

Decision Letter, first revision:

30th May 2023

Dear Dr. Keats,

I am standing in for Safia to send this decision while she is away on vacation this week; when she is back, she will be in touch with our requests (see below).

Thank you for submitting your revised manuscript "Genomic Basis of Multiple Myeloma Subtypes from the MMRF CoMMpass Study" (NG-A57979R). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

If the current version of your manuscript is in a PDF format, please email us a copy of the file in an editable format (Microsoft Word or LaTex)-- we can not proceed with PDFs at this stage.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements soon. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics. Please do not hesitate to contact me if you have any questions.

Sincerely,

Michael Fletcher, PhD
Senior Editor, Nature Genetics

ORCiD: 0000-0003-1589-7087

Reviewer #1 (Remarks to the Author):

No further comments

Reviewer #2 (Remarks to the Author):

the authors have fully addressed my concerns and I am in support of this excellent paper being published as is

Reviewer #3 (Remarks to the Author):

The authors have responded admirably to the previous critiques, and with the prolonged follow-up provide interesting and novel contributions. The CoMMpass dataset is a unique resource in myeloma as attested to the number of publications that have been based upon this data. Nevertheless, the authors are able to provide truly valuable insights with this updated analysis. I find figure 1 particularly useful as it separates genes with complete LOF, and GOF and does not simply provide a laundry list of genes. There are a couple of genes on the LOF list that I suspect do not belong there. NFKB2 is usually activated in MM, the LOF are mostly deletion and mutation, or 2 mutations. I suspect the mutations may generate constitutively activated forms of NFKB2. DIS3 is a common essential gene across almost all cancers, including myeloma. It seems unlikely there is complete LOF. Rather as described in ref 27, the mutations associated with deletions are mostly missense and appear to be only partial LOF. These are somewhat esoteric points that could perhaps be mentioned in the discussion (if that). This point certainly does invalidate this valuable analysis.

The authors have explained their definition of MYC STR. It is not clear to me why they have not included tandem duplications telomeric to MYC in the MYC STR group. The duplications involve 100-200kb centered about 300kb telomeric to MYC and account for just over 10% of MYC SV. They are analogous to the tandem duplications about 1.5kb telomeric to MYC in NOTCH1 wildtype ALL. It seems highly unlikely that their inclusion would significantly change the analysis, but it may be worth a sentence to explain their exclusion.

Author Rebuttal, first revision:

We would like to thank all three reviewers for their significant effort in reviewing this manuscript and providing valuable feedback on a study that has already had so much impact on the field.

Reviewer #1 (Remarks to the Author):

No further comments

Reviewer #2 (Remarks to the Author):

the authors have fully addressed my concerns and I am in support of this excellent paper being published as is

Reviewer #3 (Remarks to the Author):

The authors have responded admirably to the previous critiques, and with the prolonged follow-up provide interesting and novel contributions. The CoMMpass dataset is a unique resource in myeloma as attested to the number of publications that have been based upon this data. Nevertheless, the authors are able to provide truly valuable insights with this updated analysis. I find figure 1 particularly useful as it separates genes with complete LOF, and GOF and does not simply provide a laundry list of genes. There are a couple of genes on the LOF list that I suspect do not belong there. NFKB2 is usually activated in MM, the LOF are mostly deletion and mutation, or 2 mutations. I suspect the mutations may generate constitutively activated forms of NFKB2. DIS3 is a common essential gene across almost all cancers, including myeloma. It seems unlikely there is complete LOF. Rather as described in ref 27, the mutations associated with deletions are mostly missense and appear to be only partial LOF. These are somewhat esoteric points that could perhaps be mentioned in the discussion (if that). This point certainly does not invalidate this valuable analysis.

Response: This is an important point and one of the limits of the LOF analysis where it simply looks for evidence of bi-allelic loss. In the case of NFKB2 most of the inactivation events are either deletions and mutations or two or more mutations, which puts this gene in the LOF category. We did have some genes that showed up in both LOF and GOF and a process for assigning a gene to one or the other is outlined in the provided methods. In the specific case of NFKB2 there is the added complication of the two functions of the polypeptide, in its p100 isoform it is an inhibitor of NFkB signaling, but when a C-terminal truncation occurs that inhibitory, tumor suppressor like, function is lost BUT then the N-terminal portion is free to move into the nucleus and activates a pro-myeloma survival/proliferation transcriptional signature, which can be argued to be an oncogene. Unfortunately, this potential random C-terminal truncation leading to gene activation is not detectable in our GOF model so this never went through a manual assignment. In the case of DIS3 we completely agree that there is something very odd about the contribution of this gene to myeloma. One of the poorly understood phenomena is that it is frequently mutated but almost exclusively missense mutations even in those patients with deletion 13. This suggests that mutations do promote myeloma but that myeloma like seemingly all tissues also require a least one copy of the full length polypeptide. Suggesting that DIS3 has two function, one that requires the entire polypeptide that is critical to cellular survival but then a secondary function that can

be modified by a diverse set of missense mutations that promote myelomagenesis. We don't have space to address the details of NFKB2 but we added the following comment to the discussion regarding DIS3: "Oddly, DIS3 LOF events almost always maintain a full length polypeptide suggesting dual function of this gene in plasma cells"

The authors have explained their definition of MYC STR. It is not clear to me why they have not included tandem duplications telomeric to MYC in the MYC STR group. The duplications involve 100-200kb centered about 300kb telomeric to MYC and account for just over 10% of MYC SV. They are analogous to the tandem duplications about 1.5kb telomeric to MYC in NOTCH1 wildtype ALL. It seems highly unlikely that their inclusion would significantly change the analysis, but it may be worth a sentence to explain their exclusion.

Response: We do include gains that encompass MYC but these events telomeric of MYC in the PVT1 locus are not specifically included. The primary reason is that the patients with these duplications frequently also have structural rearrangements at this locus associated with the duplications and those events would be included.

Final Decision Letter:

28th Jun 2024

Dear Dr Keats,

I am delighted to say that your manuscript "Comprehensive Molecular Profiling of Multiple Myeloma Identifies Refined Copy Number and Expression Subtypes" has been accepted for publication in an upcoming issue of Nature Genetics.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Genetics style. Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

After the grant of rights is completed, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

You will not receive your proofs until the publishing agreement has been received through our system.

Due to the importance of these deadlines, we ask that you please let us know now whether you will be difficult to contact over the next month. If this is the case, we ask you provide us with the contact information (email, phone and fax) of someone who will be able to check the proofs on your behalf,

and who will be available to address any last-minute problems.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections.

You may wish to make your media relations office aware of your accepted publication, in case they consider it appropriate to organize some internal or external publicity. Once your paper has been scheduled you will receive an email confirming the publication details. This is normally 3-4 working days in advance of publication. If you need additional notice of the date and time of publication, please let the production team know when you receive the proof of your article to ensure there is sufficient time to coordinate. Further information on our embargo policies can be found here: https://www.nature.com/authors/policies/embargo.html

Before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Genetics. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact press@nature.com.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that *Nature Genetics* is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including <a href="https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative

provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we strongly recommend that you upload the step-by-step protocols used in this manuscript to protocols.io. protocols.io is an open online resource that allows researchers to share their detailed experimental know-how. All uploaded protocols are made freely available and are assigned DOIs for ease of citation. Protocols can be linked to any publications in which they are used and will be linked to from your article. You can also establish a dedicated workspace to collect all your lab Protocols. By uploading your Protocols to protocols.io, you are enabling researchers to more readily reproduce or adapt the methodology you use, as well as increasing the visibility of your protocols and papers. Upload your Protocols at https://protocols.io. Further information can be found at https://www.protocols.io/help/publish-articles.

Sincerely,

Safia Danovi, PhD
Senior Editor, Nature Genetics
ORCiD: 0009-0007-7822-5479